



The implementation of statistical arbitrage strategy on the JSE top 40 index option

JT Tsoku

 orcid.org/0000-0003-0093-6223

Thesis submitted for the degree *Doctor of Philosophy in Statistics* at
the North-West University

Promoter:

Prof ND Moroke

Graduation May 2018

Student number: 21012334



Declaration

I, Johannes Tshepiso Tsoku, hereby declare that this research report titled “The implementation of statistical arbitrage strategy on the JSE top 40 index option” is my own original work and that all sources have been accurately reported and acknowledged, and that this document has not previously in its entirety or in part been submitted at any university in order to obtain an academic qualification.



.....

J. T Tsoku

09/04/2018

.....

Date

Acknowledgements

The completion of this thesis has been possible through the help of many individuals who supported me throughout different stages of this study. First of all, I would like to thank the Almighty God for the strength He gave me, the wisdom and knowledge I needed to make this work a success. My deepest gratitude goes to the North West University for granting me the financial assistance to do my postgraduate study. I would also like to thank JSE for the data they provided.

Many thanks to my promoter, Professor Ntebogang Dinah Moroke, without her this work would not have been completed; therefore, my gratitude goes to her for her effort in making this thesis a reality. Her thoughts were always valuable. Many thanks to my mentor Daniel Metsileng, his thoughts and comments both contributed to and improved this thesis. Finally, I would also like to thank my family, friends and colleagues for their moral support and their guidance.

Abstract

This study implemented the statistical arbitrage strategy on financial time series. This strategy is known as an extended horizon trading prospect used for generating profit that is riskless and is further designed for exploiting tenacious incongruities. The study used statistical arbitrage strategy to explore its efficiency on selected stock markets in pricing relative risk in highly correlated and cointegrated stock options traded on the Johannesburg Stock Exchange (JSE). Statistical arbitrage avoids the dual premise dilemma associated with traditional tests of market efficacy. This is due to autonomous description of model symmetry and the fact that the said strategy has a mismatch with market competency. The study explored the methodology for investigating statistical arbitrage and provided the findings that determined whether or not value trading policies constitute opportunities for statistical arbitrage.

The data used are related to four sectors: the financial, industrial, mining and retail that trade on the JSE. These data are for stock market comprising closing prices of option stock for 40 companies of the four sectors traded on the JSE. The data was collected daily from 04 January 2010 to 31 December 2015. A total of 1500 observations with one year approximately having 250 trading days excluding weekends and public holidays was used.

The study followed the methodology of two broad components: simple rolling regression and Engle and Granger cointegration technique. A two-step approach, namely correlation analysis and cointegration technique, were used in formulating potential pairs portfolio on the basis of the profitability indicator evaluated in-sample. The findings from correlational analysis revealed thirty nine pairs under financial sector, two pairs under industrial sector, five pairs under mining sector and six pairs under retail sector. The selected pairs were stationary at first log difference. The findings of cointegration analysis confirmed only eighteen pairs under financial sector, one pair in industrial sector, two pairs in mining sector and three pairs in retail sector. Since the results revealed a mean reversion, statistical arbitration strategy was implemented. The findings revealed that the financial sector performed better than the other three sectors. The implementation of pairs trading to historical volatility revealed several possibilities of trade in the financial sector. Therefore, the investors profit more from trading the identified pairs of stocks in the financial sector.

The study contributes to the existing literature in historical volatility modelling and stochastic volatility in the field of index option in emerging markets. This study further provided a base for future researchers conducting studies on emerging markets, more specifically in South African context.

Table of Contents

Declaration.....	i
Dedication.....	ii
Acknowledgements.....	iii
Abstract.....	iv
Table of Contents.....	vi
List of Acronyms.....	xvi
CHAPTER 1.....	1
ORIENTATION OF THE STUDY.....	1
1.1 INTRODUCTION.....	1
1.2 BACKGROUND TO THE STUDY.....	2
1.3 PROBLEM STATEMENT.....	4
1.4 RATIONALE OF THE STUDY.....	5
1.5 AIM AND OBJECTIVES.....	6
1.6 RESEARCH QUESTIONS.....	6
1.7 SIGNIFICANCE OF THE STUDY.....	6
1.8 DELIMITATIONS OF THE STUDY.....	7
1.9 DEFINITION OF TERMS.....	7
1.10 RESEARCH OUTLINE.....	8
1.11 CHAPTER SUMMARY.....	8
CHAPTER 2.....	9
THEORETICAL FRAMEWORK AND LITERATURE REVIEW.....	9
2.1 INTRODUCTION.....	9
2.2 THEORETICAL FRAMEWORK.....	9
2.2.1 Characteristics of financial time series data.....	9
2.2.2 Historical Volatility.....	10
2.2.3 Stationarity.....	11
2.2.4 Statistical arbitrage.....	14
2.2.5 Pair trading strategy.....	16
2.2.6 Cointegration.....	19
2.2.7 Distance trading method.....	22
2.2.8 Stochastic spread method.....	23

2.2.9	The combined forecast method	24
2.3	EMPIRICAL LITERATURE.....	25
2.4	CHAPTER SUMMARY	33
CHAPTER 3		35
RESEARCH METHODOLOGY		35
3.1	INTRODUCTION.....	35
3.2	ETHICAL CONSIDERATIONS	35
3.3	RESEARCH PROCESS.....	36
3.3.1	Philosophy.....	36
3.3.2	Research approach.....	37
3.3.3	Research strategy.....	38
3.3.4	Choice of research	38
3.4	IDENTIFYING NUMBER OF PAIRS.....	40
3.5	DATA ANALYSIS FRAMEWORK	41
3.6	PRELIMINARY ANALYSIS	41
3.7	SELECTION OF INDEXES	42
3.7.1	Correlation analysis.....	42
3.7.2	Data transformation.....	44
3.8	TESTS FOR STATIONARITY	45
3.8.1	Augmented Dickey-Fuller (ADF) Test	46
3.8.2	Kwiatkowski, Phillips, Schmidt and Shin (KPSS) Test.....	47
3.9	RELATIONSHIP BETWEEN INDEX PAIRS	48
3.10	BOUNDARIES FOR REGRESSION COEFFICIENTS.....	50
3.11	PAIRS TRADING USING COINTEGRATION TECHNIQUE.....	50
3.11.1	Verification of cointegration	52
3.12	ERROR CORRECTION MODEL (ECM)	54
3.13	MODEL DIAGNOSTICS	55
3.13.1	Serial correlation test.....	56
3.13.2	Heteroscedasticity test.....	57
3.13.3	Normality test.....	58

3.14	CAUSALITY TEST.....	59
3.15	STATISTICAL ARBITRAGE TRADING STRATEGY.....	60
3.16	CHAPTER SUMMARY.....	63
CHAPTER 4.....		64
DATA ANALYSIS AND INTERPRETATION OF THE RESULTS.....		64
4.1	INTRODUCTION.....	64
4.2	RELIMINARY DATA ANALYSIS.....	64
4.2.1	Analysis on financial sector.....	64
4.2.2	Graphical presentation of financial sector.....	65
4.2.3	Test for stationarity.....	68
4.2.4	Relationship between the index pairs.....	70
4.2.5	Verification of cointegration.....	85
4.2.6	Error Correction Model.....	86
4.2.7	Model diagnostics results.....	88
4.2.8	Granger Causality results.....	91
4.2.9	Implementing pairs trading strategy.....	92
4.3	PRELIMINARY DATA ANALYSIS.....	99
4.3.1	Analysis on industrial sector.....	99
4.3.2	Graphical presentation of industrial sector.....	100
4.3.3	Test for stationarity.....	102
4.3.4	Relationship between the index pairs.....	103
4.3.5	Verification of cointegration.....	104
4.3.6	Error Correction Model.....	105
4.3.7	Model diagnostics.....	106
4.3.8	Granger Causality results.....	106
4.3.9	Implementing pairs trading strategy.....	107
4.4	PRELIMINARY DATA ANALYSIS.....	108
4.4.1	Analysis on mining sector.....	108
4.4.2	Graphical presentation of mining sector.....	108

4.4.3	Test for stationarity	111
4.4.4	Relationship between the index pairs.....	112
4.4.5	Verification of Cointegration	115
4.4.6	Error Correction Model.....	116
4.4.7	Model diagnostics	117
4.4.8	Granger Causality results	118
4.4.9	Implementing pairs trading strategy.....	118
4.5	PRELIMINARY DATA ANALYSIS.....	119
4.5.1	Analysis on retail sector	119
4.5.2	Graphical presentation of retail sector	120
4.5.3	Test for stationarity	123
4.5.4	Relationship between the index pairs.....	124
4.5.5	Verification of Cointegration	127
4.5.6	Error Correction Model.....	128
4.5.7	Model diagnostics	129
4.5.8	Granger Causality results	130
4.5.9	Implementing pairs trading strategy.....	130
4.6	CHAPTER SUMMARY	132
CHAPTER 5		134
CONCLUSION AND RECOMMENDATIONS		134
5.1	INTRODUCTION.....	134
5.2	DISCUSSION OF RESULTS.....	134
5.3	CONTRIBUTION TO THE STUDY	139
5.4	LIMITATIONS	140
5.5	CONCLUSION	141
5.6	RECOMMENDATIONS FOR FURTHER STUDIES.....	141
5.7	SUMMARY OF THE THESIS.....	142
REFERENCES		143
APPENDICES		157

List of Figures

Figure 2.1: Graphical presentation of pairs trading strategy (Yakop, 2011)	18
Figure 3.1: Research onion by Saunders et al. (2012)	36
Figure 4.1.1: Time series plot of the historical volatility of financial sector at level	66
Figure 4.1.2: Time series plot of the financial sector at logarithm	66
Figure 4.1.3: Time series plot of the financial sector at logarithm difference	67
Figure 4.1.4: Beta boundaries for $\Delta LDSY-\Delta LBVT$, $\Delta LFSR-\Delta LBVT$, $\Delta LINL-\Delta LBVT$, $\Delta LNED-\Delta LBVT$, $\Delta LOML-\Delta LBVT$, $\Delta LREI-\Delta LBVT$, $\Delta LRMH-\Delta LBVT$ and $\Delta LSLM-\Delta LBVT$	71
Figure 4.1.5: Beta boundaries for $\Delta LFSR-\Delta LDSY$, $\Delta LINL-\Delta LDSY$, $\Delta LNED-\Delta LDSY$, $\Delta LOML-\Delta LDSY$, $\Delta LREI-\Delta LDSY$, $\Delta LRMH-\Delta LDSY$, $\Delta LSLM-\Delta LDSY$ and $\Delta LINL-\Delta LFSR$	74
Figure 4.1.6: Beta boundaries for $\Delta LNED-\Delta LFSR$, $\Delta LOML-\Delta LFSR$, $\Delta LREI-\Delta LFSR$, $\Delta LRMH-\Delta LFSR$, $\Delta LSBK-\Delta LFSR$, $\Delta LSLM-\Delta LFSR$, $\Delta LOML-\Delta LINL$ and $\Delta LREI-\Delta LINL$	77
Figure 4.1.7: Beta boundaries for $\Delta LRMH-\Delta LINL$, $\Delta LSLM-\Delta LINL$, $\Delta LOML-\Delta LNED$, $\Delta LREI-\Delta LNED$, $\Delta LRMH-\Delta LNED$, $\Delta LSBK-\Delta LNED$, $\Delta LSLM-\Delta LNED$ and $\Delta LREI-\Delta LOML$	80
Figure 4.1.8: Beta boundaries for $\Delta LRMH-\Delta LOML$, $\Delta LSLM-\Delta LOML$, $\Delta LRMH-\Delta LREI$, $\Delta LSLM-\Delta LREI$, $\Delta LSBK-\Delta LRMH$, $\Delta LSLM-\Delta LRMH$ and $\Delta LSLM-\Delta LSBK$	83
Figure 4.1.9: Historical movement of $LDSY-LBVT$, $LOML-LBVT$, $LSLM-LBVT$, $LFSR-LBVT$, $LREI-LBVT$ and $LFSR-LDSY$ pairs	93
Figure 4.1.10: Historical movement of $LOML-LDSY$, $LREI-LDSY$, $LOML-LFSR$, $LREI-LFSR$, $LSLM-LFSR$ and $LNED-LFSR$ pairs	94
Figure 4.1.11: Historical movement of $LRMH-LINL$, $LSLM-LNED$, $LREI-LOML$, $LSLM-LOML$, $LSLM-LREI$ AND $LSBK-LRMH$ pairs	95
Figure 4.1.12: Spread of $LDSY-LBVT$, $LOML-LBVT$, $LSLM-LBVT$, $LFSR-LBVT$, $LREI-LBVT$ and $LFSR-LDSY$ pairs with trading thresholds	96
Figure 4.1.13: Spread of $LOML-LDSY$, $LREI-LDSY$, $LOML-LFSR$, $LREI-LFSR$, $LSLM-LFSR$ and $LNED-LFSR$ pairs with trading thresholds	97

Figure 4.1.14: Spread of LRMH-LINL, LSLM-LOML, LSLM-LNED, LREI-LOML, LSLM-LREI and LSBK-LRMH pairs with trading thresholds	98
Figure 4.2.1: Time series plot of the historical volatility of industrial sector at level	100
Figure 4.2.2: Time series plot of the historical volatility of industrial sector at logarithm ...	101
Figure 4.2.3: Time series plot of the historical volatility of industrial sector at logarithm difference	101
Figure 4.2.4: Beta boundaries for $\Delta\text{LNMK} - \Delta\text{LMTN}$ and $\Delta\text{LNTC} - \Delta\text{LMND}$	103
Figure 4.2.5: Historical movement of LNMK – LMTN pair	107
Figure 4.2.6: The residual series of LNMK – LNMK pair	107
Figure 4.3.1: Time series plot of the historical volatility of mining sector at level	109
Figure 4.3.2: Time series plot of the historical volatility of mining sector at logarithm	109
Figure 4.3.3: Time series plot of the historical volatility of mining sector at logarithm difference	110
Figure 4.3.4: Beta boundaries for $\Delta\text{LAMS} - \Delta\text{LACL}$ and $\Delta\text{LIMP} - \Delta\text{LACL}$	112
Figure 4.3.5: Beta boundaries for $\Delta\text{LIMP} - \Delta\text{LAGL}$ and $\Delta\text{LIMP} - \Delta\text{LAMS}$	113
Figure 4.3.6: Beta boundaries for $\Delta\text{LGFI} - \Delta\text{LANG}$	114
Figure 4.3.7: Historical movement of LAMS – LACL pair and LIMP – LACL pair	118
Figure 4.3.8: The residual series of LAMS – LACL and LIMP – LACL pair	119
Figure 4.4.1: Time series plot of the historical volatility of retail sector at level	121
Figure 4.4.2: Time series plot of the historical volatility of retail sector at logarithm	121
Figure 4.4.3: Time series plot of the historical volatility of retail sector at logarithm difference	122
Figure 4.4.4: Beta boundaries for $\Delta\text{LBTI} - \Delta\text{LAPN}$ and $\Delta\text{LSAB} - \Delta\text{LAPN}$	124
Figure 4.4.5: Beta boundaries for $\Delta\text{LSAB} - \Delta\text{LBTI}$ and $\Delta\text{LWHL} - \Delta\text{LBTI}$	125
Figure 4.4.6: Beta boundaries for $\Delta\text{LWHL} - \Delta\text{LSAB}$ and $\Delta\text{LWHL} - \Delta\text{LTSG}$	127
Figure 4.4.7: Historical movement of LSAB – LBTI and LWHL – LBTI pair	130
Figure 4.4.8: Historical movement of LWHL – LSAB pair	131

Figure 4.4.9: The residual series of LSAB – LBTI and LWHL – LBTI pair 131

Figure 4.4.10: The residual series of LWHL – LSAB pair..... 132

List of Tables

Table 3.1: Stock market data	40
Table 3.2: Analysis framework.....	41
Table 4.1.1: Descriptive statistics for log transformed financial sector	65
Table 4.1.2: Pairwise correlation of the financial sector	68
Table 4.1.3: Unit root tests of the financial sector	68
Table 4.1.4: Beta boundaries for $\Delta LDSY-\Delta LBVT$, $\Delta LFSR-\Delta LBVT$, $\Delta LINL-\Delta LBVT$, $\Delta LNED-\Delta LBVT$, $\Delta LOML-\Delta LBVT$, $\Delta LREI-\Delta LBVT$, $\Delta LRMH-\Delta LBVT$ and $\Delta LSLM-\Delta LBVT$	72
Table 4.1.5: Beta boundaries for $\Delta LFSR-\Delta LDSY$, $\Delta LINL-\Delta LDSY$, $\Delta LNED-\Delta LDSY$, $\Delta LOML-\Delta LDSY$, $\Delta LREI-\Delta LDSY$, $\Delta LRMH-\Delta LDSY$, $\Delta LSLM-\Delta LDSY$ and $\Delta LINL-\Delta LFSR$	75
Table 4.1.6: Beta boundaries for $\Delta LNED-\Delta LFSR$, $\Delta LOML-\Delta LFSR$, $\Delta LREI-\Delta LFSR$, $\Delta LRMH-\Delta LFSR$, $\Delta LSBK-\Delta LFSR$, $\Delta LSLM-\Delta LFSR$, $\Delta LOML-\Delta LINL$ and $\Delta LREI-\Delta LINL$	78
Table 4.1.7: Beta boundaries for $\Delta LRMH-\Delta LINL$, $\Delta LSLM-\Delta LINL$, $\Delta LOML-\Delta LNED$, $\Delta LREI-\Delta LNED$, $\Delta LRMH-\Delta LNED$, $\Delta LSBK-\Delta LNED$, $\Delta LSLM-\Delta LNED$ and $\Delta LREI-\Delta LOML$	81
Table 4.1.8: Beta boundaries for $\Delta LRMH-\Delta LOML$, $\Delta LSLM-\Delta LOML$, $\Delta LRMH-\Delta LREI$, $\Delta LSLM-\Delta LREI$, $\Delta LSBK-\Delta LRMH$, $\Delta LSLM-\Delta LRMH$ and $\Delta LSLM-\Delta LSBK$	84
Table 4.1.9: Engle and Granger results for 39 paired indices in financial sector	85
Table 4.1.10: ECM results for the cointegrated pairs of the financial sector	87
Table 4.1.11.1: Diagnostic test results for $LDSY-LBVT$, $LFSR-LBVT$, $LOML-LBVT$, $LREI-LBVT$, $LSLM-LBVT$, $LFSR-LDSY$, $LOML-LDSY$ and $LREI-LDSY$	88
Table 4.1.11.2: Diagnostic test results for $LNED-LFSR$, $LOML-LFSR$, $LREI-LFSR$, $LSLM-LFSR$, $LRMH-LINL$, $LSLM-LNED$, $LREI-LOML$ and $LSLM-LOML$	89
Table 4.1.11.3: Diagnostic test results for $LSLM-LREI$ and $LSBK-LRMH$	90
Table 4.1.12.1: Causality results for $LBVT-LDSY$, $LBVT-LFSR$, $LBVT-LOML$, $LBVT-LREI$, $LBVT-LSLM$, $LDSY-LFSR$, $LDSY-LOML$, $LDSY-LREI$ and $LFSR-LNED$	91
Table 4.1.12.2: Causality results for $LFSR-LOML$, $LFSR-LREI$, $LFSR-LSLM$, $LINL-LRMH$, $LNED-LSLM$, $LOML-LREI$, $LOML-LSLM$, $LREI-LSLM$ and $LRMH-LSBK$	92

Table 4.2.1: Descriptive statistics for log transformed industrial sector	99
Table 4.2.2: Pairwise correlation of the industrial sector	102
Table 4.2.3: Unit root tests results for the industrial sector	102
Table 4.2.4: Beta boundaries for $\Delta\text{LNMK} - \Delta\text{LMTN}$ and $\Delta\text{LNTC} - \Delta\text{LMND}$	104
Table 4.2.5: Engle and Granger cointegration results for $\text{LNMK} - \text{LMTN}$ and $\text{LNTC} - \text{LMND}$	105
Table 4.2.6: ECM results for LNMK-LMTN	105
Table 4.2.7: Diagnostic results for LNMK-LMTN	106
Table 4.2.8: Causality between LMTN and LNMK	106
Table 4.3.1: Descriptive statistics of log transformed mining sector	108
Table 4.3.2: Pairwise correlation of the mining sector	110
Table 4.3.3: Unit root tests of the mining sector.....	111
Table 4.3.4: Beta boundaries for $\Delta\text{LAMS} - \Delta\text{LACL}$ and $\Delta\text{LIMP} - \Delta\text{LACL}$	112
Table 4.3.5: Beta boundaries for $\Delta\text{LIMP} - \Delta\text{LAGL}$ and $\Delta\text{LIMP} - \Delta\text{LAMS}$	114
Table 4.3.6: Beta boundaries for $\Delta\text{LGFI} - \Delta\text{LANG}$	115
Table 4.3.7: Engle and Granger cointegration results for $\text{LAMS} - \text{LACL}$, $\text{LIMP} - \text{LACL}$, $\text{LIMP} - \text{LAGL}$, $\text{LIMP} - \text{LAMS}$ and $\text{LGFI} - \text{LANG}$	116
Table 4.3.8: ECM results for $\text{LAMS} - \text{LACL}$ and $\text{LIMP} - \text{LACL}$	116
Table 4.3.9: Diagnostic results for $\text{LAMS} - \text{LACL}$ and $\text{LIMP} - \text{LACL}$	117
Table 4.3.10: Causality between $\text{LAMS} - \text{LACL}$ and $\text{LIMP} - \text{LACL}$	118
Table 4.4.1: Descriptive statistics of log transformed retail sector.....	120
Table 4.4.2: Pairwise correlation of the retail sector	122
Table 4.4.3: Unit root tests of the retail sector.....	123
Table 4.4.4: Beta boundaries for $\Delta\text{LBTI} - \Delta\text{LAPN}$ and $\Delta\text{LSAB} - \Delta\text{LAPN}$	124
Table 4.4.5: Beta boundaries for $\Delta\text{LSAB} - \Delta\text{LBTI}$ and $\Delta\text{LWHL} - \Delta\text{LBTI}$	126
Table 4.4.6: Beta boundaries for $\Delta\text{LWHL} - \Delta\text{LSAB}$ and $\Delta\text{LWHL} - \Delta\text{LTSG}$	127

Table 4.4.7: Engle and Granger cointegration results for LBTI – LAPN, LSAB – LAPN, LSAB – LBTI, LWHL – LBTI, LWHL – LSAB and LWHL – LTSG 128

Table 4.4.8: ECM results for LSAB – LBTI, LWHL – LBTI and LWHL – LSAB 128

Table 4.4.9: Diagnostic results for LSAB – LBTI, LWHL – LBTI and LWHL – LSAB..... 129

Table 4.4.10: Causality between LSAB – LBTI, LWHL – LBTI and LWHL – LSAB..... 130

List of Acronyms

ACL	Arcelor Mittal South Africa Ltd
ADF	Augmented Dickey-Fuller
AGL	Anglo American
AIC	Akaike Information Criterion
AltX	Alternative Exchange
AMS	Anglo American Platinum
ANG	AngloGold
APN	Aspen Pharmacare Holdings
ARCH	Autoregressive Conditional Heteroscedasticity
ARI	African Rainbow Minerals Gold
ARIMA	Autoregressive Integrated Moving Average
BESA	Bond Exchange of South Africa
BG	Breusch-Godfrey
BGA	Barclays Africa Group Ltd
BIC	Bayesian Information Criterion
BIL	BHP Billiton
BTI	British American Tobacco PLC
BVT	Bidvest Group
DSY	Discovery Ltd
ECM	Error Correction Model
ETFs	Exchange Traded Funds
FPE	Final Prediction Error
FSR	FirstRand Limited
GARCH	Generalised Autoregressive Conditional Heteroscedasticity
GFI	Gold Fields Limited
HQC	Hannan-Quinn Criterion

IMP	Impala Platinum Holdings
INL	Investec Limited
IPL	Imperial Holdings
JB	Jarque-Bera
JSE	Johannesburg Stock Exchange
KIO	Kumba Iron Ore
KPSS	Kwiatkowski-Phillips-Schmidt-Shin
LB	Ljung-Box
LM	Lagrange Multiplier
MND	Mondi Ltd
MTN	MTN Group
NED	Nedbank Group
NMK	Nampak
NTC	Netcare Limited
NYSE	New York Stock Exchange
OLS	Ordinary Least Squares
OML	Old Mutual
PCA	Principal Component Analysis
PIK	Pick'n Pay Stores
PP	Phillips-Perron
PPC	Pretoria Portland Cement
REI	Reinet Investments
RMH	RMB Holdings
SAB	SABMiller
SAFEX	South African Futures Exchange
SAP	Sappi
SBC	Schwarz Bayesian Criterion

SBK	Standard Bank Group
SD	Standard Deviation
SHP	Shoprite
SLM	Sanlam
SOL	Sasol
TBS	Tiger Brands
TFG	The Foschini Group
TKG	Telkom
TRU	Truworths International
TSG	The Spar Group
US	United States
VAR	Vector Autoregressive
VECM	Vector Error Correction Model
VOD	Vodacom Group
WHL	Woolworths Holdings

CHAPTER 1

ORIENTATION OF THE STUDY

1.1 INTRODUCTION

Financial derivatives play a very significant role in any economic system since they allow investors to hedge their risks and to increase their returns (Kiiashko, 2012). These financial derivatives are important tools for investors to manage risk in the current financial world. A derivative is an instrument whose value depends on the value of other variables such as indices, stocks, interest rates, treasury bills, and so on. In other words, the value of a derivative depends on the value of the underlying asset (Wilmott, 1998). There are high possibilities of making profit when individuals and companies are knowledgeable about stock markets. There are also many kinds of financial derivatives such as futures or options, forwards and swamps. However, the present study only focuses on option pricing. The stock market has attracted the attention of both individuals and companies as practitioners and academics seek to find alternative ways of protecting financial commodities from unfavourable market conditions. The interest in the stock market is primarily based on the probability of minimising risk and maximising profits.

Options are derivatives that give the buyer the **right to buy or sell a predetermined (strike) quantity of the underlying asset from or to the writer of the option in a predetermined time in the future (maturity time)** (Garman and Forgue, 2015). The writer then has the right to sell to or buy the underlying asset. There are two categories of options, namely, calls¹ and puts². Call options give the right to the holder (buyer) of the option to **buy the predetermined quantity of the underlying asset from its writer, meaning that the writer has the obligation to sell the underlying asset**. Put options give the right to the seller to **sell a predetermined quantity of the underlying asset to the writer of the option** (Hull, 1997; Wilmott, 1998; Garman and Forgue, 2012). Options give market participants **great leverage by reducing the risk (hedging) and arbitrage opportunities**. They also give participants **the opportunity to reduce losses while maintaining the possibility of profiting from favourable changes in the underlying asset**. These opportunities can be exploited only if the options are combined with other variables such as stocks, indices or interest rates, or if they are combined. These two kinds of combinations are called option strategies (Syz, 2008). Volatility works its way through every strategy. The

¹ Calls increase in value when the underlying security is going up and they decrease in value when the price of the underlying security declines.

² Puts increase in value when the underlying security is going down and decrease in value when it is going up.

interest on options continue to grow and this leads to markets being volatile and also affects the pricing of options (Investopedia, 2015).

Volatility is defined by Schwert (1990) as a statistical measure of dispersion around a mean value. It is also defined under a theoretical aspect as the changeability or randomness of the underlying asset. This measure of volatility is usually viewed from three different perspectives: historical (backward looking), implied (reflected in an option market price) or future (forward looking). Historical volatility involves computing the variance or standard deviation of stock returns in the usual way over some historical period, and this then becomes the volatility forecast for all periods in the future. Volatility can be modelled using the ordinary least squares (OLS), autoregressive conditional heteroscedasticity (ARCH) and generalised autoregressive conditional heteroscedasticity (GARCH) specifications.

De Gooijer and Hyndma (2006) posit that over the years, financial data have been analysed using statistical and econometric techniques. Chatfield (2004) highlighted that classical time series models such as ARCH, autoregressive integrated moving average (ARIMA), GARCH and many other extensions and variations are frequently used to explore the financial data and make predictions. The author further states that conventional time series and other forecasting models may not address challenges faced when developing modern algorithmic trading strategies. These techniques are mostly exploratory in nature and they are intended to detect the dataset patterns that may be continuously changing and evolving over time. Therefore, OLS regression may be used to cater for time-varying dependencies between the paired streams of data. In order to allow for time-variant regression coefficient, Tesfatsion and Kalaba (1989) introduced the flexible least square as a generalization of the standard linear regression model. Flexible regression also follows the assumptions of usual regression. In order to solve the problem of time variation, the current study models historical volatility using the flexible OLS approach which is also called the rolling OLS approach and cointegration approach.

1.2 BACKGROUND TO THE STUDY

The Johannesburg Stock Exchange (JSE) was formed in 1887 during the first South African gold rush. In 1963, the JSE joined the World Federation of Exchanges following the first legislation that covered financial markets in 1947 and upgraded to an electronic trading system in the early 1990s (Oxford Business Group, 2013). Prior to 1994, the JSE faced challenges

including a rigid exchange control regime, international economic sanctions, thin trading and illiquidity, political instability and peculiar economic structures of companies.

The trading of financial institutions on the JSE increased significantly subsequent to the constant cash flow from insurance and pension funds over the years. Investors encountered challenges in achieving smooth and efficient trade in shares. All these challenges explain why the JSE performed poorly before 1994 when compared to other emerging markets (Ndako, 2010). Therefore, the statistical arbitrage technique seeks to profit from the statistical mispricing of one or more assets based on the expected value of the assets generated from a statistical model.

In the year 2003, the JSE launched an alternative exchange (AltX) for small and mid-sized listings, followed by the Yield X for currency instruments and interest rates. In 2001, the JSE then acquired the South African Futures Exchange (SAFEX) and the Bond Exchange of South Africa (BESA) later in 2009. Also in 2007, the JSE launched currency derivatives which provide opportunities for market participants to hedge against currency risk, and diversify internationally as well as taking a view on the movement of the underlying exchange rates. To date, the JSE offers five financial markets, namely, bonds, equities, financial, commodity and interest rate derivatives. The JSE offers trading of a diversity of derivatives, including futures and options on equities, indices, bonds, currencies, commodities and interest rates.

In 2012, the JSE was ranked the 6th largest exchange by number of Single Stock Futures traded and 9th by the number of currency derivatives traded in the World Federation of Exchanges Annual Derivatives Market Survey. South Africa is currently ranked 1st in the world in terms of regulation of securities exchanges in the World Economic Forum's Global Competitiveness Survey for 2013/2014 (ASEA, 2015). The JSE is indexed by the FTSE/JSE Africa Index Series, a partnership between JSE and the FTSE group (Oxford Business Group, 2013). The construct of option has been used since ancient times. The first well-known option market was established in Holland in the early 17th century, trading tulips. At the beginning, the purpose of options was to serve as security for producers and merchandisers, but later the market attracted many speculators. This incidence, known as the tulip mania, is the first recorded economic bubble in history (Garber, 1990).

The rating agencies such as Moody's and Standard and Poor have rated the economy of South Africa as one of the most developing economies in a developing country. In November 2014, the ratings agencies Fitch and Standard and Poor granted South Africa a stay of execution, keeping its credit rating unchanged. However, Moody's downgraded it one notch to BAA2, just two above "junk status". The CIA World Factbook (2015) indicated that inflation, unemployment, unstable electricity supply and instability in the currency are factors that negatively affect the economy of South African and its financial markets. More specifically, traders in the South African financial markets often refer to the lack of liquidity as a cause of the inefficiency of the market. The economies of the emerging market, which are going through a period of economic instability, may therefore be the cause of instability in the South African economy.

1.3 PROBLEM STATEMENT

The pairs traders/investors most of the time find it difficult to decide when to trade and when to hold stock, therefore the results of statistical arbitrage are important as they help them in making a financially sound decision. The more risky types of arbitrage, such as statistical and volatility arbitrage, have received more attention from an international perspective but this attention is not evident in the context of South Africa. Despite their high practical relevance, these types of arbitrage have also received little attention. Some authors point out to a stumpy application of these methods as a result of associated mispricing. There is a need for studies that use statistical arbitrage as an innovative method to analyse financial time series as these methods are based on statistically established relationships rather than theoretical exact pricing relations. There is also a need for studies that could help in understanding market mechanisms and market efficiency such as the current one.

Literature highlights on the drawbacks displayed by the traditional methodologies used in previous studies in the field of stock markets. These methodologies suffer a deficiency of integrating modelling framework that concerns among others; the process of pre-processing of data, predictive modelling and the procedure that implements the decision obtained from the second step. These three steps are of utmost importance and pose a disadvantage if not implemented due to the fact that in practice, any one of these steps is dependent upon statistical mispricings and predictive information. Hence, this study implements an innovative strategy involving the three steps mentioned on the Top 40 index options to try and fill a gap identified and also to set a platform for other scholars who are interested in this area.

The Competition Commission of South Africa has had an influx of cases relating to price instability in the country. The financial market in South Africa is characterised as one of the emerging markets within the world economy. World trends depict that emerging markets tend to provide excess returns through market inefficiencies. This may be exploited through the operations of the statistical arbitrage techniques. The key feature to most of the inefficiencies has seemingly been the pricing factor. These opportunities have remained topical since the late 1980s. This is reflected in an article in the Economist of 21 July 1990 and is still topical. Statistical arbitrage is also visible in current studies such as one by Ammann and Herriger (2002). The current study is planned around similar scope and framework in the context of South Africa. Price fluctuations in a regulated environment are permissible if it is done across the board to guard against unfair advantage and competition.

Market inefficiencies that are experienced in the international context are even more experienced in the South African setting where statistical arbitrage operations are frequently completely removed from one another. The method used to manage the profitability, trading opportunities and risks are even more unstructured due to the dispersion of the functions. Studies on the subject reveal that market inefficiencies are decreasing over time provided that the depository receipts, restriction on short sale and transaction costs are kept minimal. This study lays a foundation for researchers in the field and also encourages the application of statistical arbitrage as far as JSE Top 40 Index Option is concerned. Specifically, the study fills a gap by implementing statistical arbitrage pairs trading strategy to stock prices that are highly correlated and cointegrated in South African context. The study furthermore seeks to test where trade is possible in the four different sectors trading on JSE through statistical arbitrage method.

1.4 RATIONALE OF THE STUDY

The main goal of this study is to apply a statistical arbitrage strategy to investigate the efficiency of stock markets in pricing relative risk in highly correlated stocks options traded on JSE. This study provides a base for future researchers conducting studies on emerging markets, more specifically in South African context. The application of statistical arbitrage trading strategy in the South African context is very limited, therefore this study paves a way in the application of the said technique.

1.5 AIM AND OBJECTIVES

The study aims to develop a modelling framework for testing the efficiency of stock markets in pricing relative risk in highly correlated stock options traded on JSE.

The objectives of the study are set to:

- identify pairs of highly correlated index options,
- determine whether or not there is a long run relationship (cointegration) between the selected stock indexes,
- determine the direction(s) of causality between the paired stock indexes,
- identify which pairs of stock are to be traded,
- use the findings of the study in formulating suggestions for investor decision-making purposes.

1.6 RESEARCH QUESTIONS

The research questions of the study are as follows:

- Which pairs from the Top 40 index options are highly correlated?
- Which pairs of stock are to be traded?
- Is there a long run relationship (cointegration) between the selected stock indexes?
- What is the direction of causality between the paired stock indexes?
- What suggestions could be formulated for investors and portfolio managers?

1.7 SIGNIFICANCE OF THE STUDY

The findings of this study might provide direction for future research seeking to advance the forecast quality of historical market volatility. Moreover, the findings could be useful in improving the performance of option valuation models. The study strives to benefit scholars who are doing research in the area for the application of statistical arbitrage to index option. Portfolio managers may benefit from the study when they refer to it in making investment practice and index option pricing decisions. Also, practitioners including the investors and company executives, who are in search of a better understanding of index option pricing, may find this study helpful. Furthermore, the findings of this study may be beneficial to participants trading on the JSE and South African government monetary policy makers.

1.8 DELIMITATIONS OF THE STUDY

The study confines itself only to modelling historical volatility by using statistical arbitrage technique in a South African perspective. The data used in this study is secondary and is closing prices of option and stock for 40 companies traded on JSE. The data was collated from 04 January 2010 to 31 December 2015. No surveys, interviews or questionnaires are conducted in this study. The researcher also did not cover other financial variables other than the ones used in the study. The findings of this study may not be generalised as they only apply to the variables used in the study within the South African context.

1.9 DEFINITION OF TERMS

- Arbitrage - is the profit making market activity of buying and selling of same stock on different exchanges (stock market) or between spot prices of a stock and its future contract (Investopedia, 2015).
- Volatility - is a statistical measure of dispersion around a mean value; it is the changeability or randomness of the underlying asset (Schwert, 1990).
- Derivative - is a security whose price is dependent upon or derived from one or more underlying assets (Investopedia, 2015).
- Liquidity - is the degree to which an asset or security can be bought or sold in the market without affecting the stock price (Investopedia, 2015).
- Mispricing - The quality of a security, good or service having a price which does not correctly match the intrinsic value of the item (Marwan, 2015).
- Economic bubble - is a surge in the market caused by speculation regarding a commodity which results in an explosion of activity in that market segment causing vastly overinflated prices (Business Dictionary, 2015).
- Unit Root/stationary - A series is referred to as “(weakly or covariance) stationary if its mean and variance are constant over time and the value of the covariance between the two time periods depends only on the distance or lag between the two time periods, not on the time at which the covariance is calculated” (Gujarati, 2003).
- Hedge - Making an investment to reduce the risk of adverse price movement in an asset. Investors use this strategy when they are not sure of what the market is likely to do (Investopedia, 2015).

1.10 RESEARCH OUTLINE

Chapter 1 outlines the introduction and background to the study, problem statement, aim and objectives of the study. Chapter 1 also outlines the research objectives and questions, purpose of the research, research methodology, contribution of the study, brief literature review, and scope the of the research, together with limitations/delimitation of the study and definition of terms.

Chapter 2 reviews relevant and recent literature. This chapter also provides a review of finance literature with a focus on key concepts and issues in options/equity markets and statistical arbitrage.

Chapter 3 discusses the methodology of the study. The research procedure comprises of the ethical considerations, research process, and execution. The chapter also provides a research framework as well as test for stationarity using the ADF and KPSS unit root test, OLS regression, cointegration and statistical arbitrage trading strategy.

Chapter 4 provides a detailed statistical analysis and interpretation of the results generated from the dataset.

Lastly, Chapter 5 provides the conclusion and recommendations.

1.11 CHAPTER SUMMARY

This chapter provided the context and background to the research problem identified, the orientation to the research context, and the structure and plan of the research. Research objectives were clearly aligned to the title and the research problem. Furthermore, the chapter highlighted some of the potential benefits of the study and the boundaries the study was confined to.

The next chapter provides a critical review of literature related to the study and some of the important theories that relate to stock exchanges and statistical arbitrage techniques.

CHAPTER 2

THEORETICAL FRAMEWORK AND LITERATURE REVIEW

2.1 INTRODUCTION

This chapter reviews the empirical literature of studies conducted in this field and sets the theoretical framework for the application of the statistical arbitrage method on JSE Top 40 index option. The intention is to identify knowledge gaps in this literature on the study of statistical arbitrage. Most of the studies reviewed were conducted internationally and there is a paucity of such studies in South Africa currently. The first section of this chapter focuses on the theoretical framework and the second one focuses on the empirical literature.

2.2 THEORETICAL FRAMEWORK

This section presents the theoretical framework, constructs and domain definitions of statistical arbitrage and analysis of financial data. Maier (2013) suggested that the initial step in identifying and articulating knowledge gaps in literature reviews is to identify the domain problem. This establishes the gaps that a particular study seeks to bridge and therefore make a contribution. For the current study, the problem identified is the applicability and efficiency of statistical arbitrage method to the South African closing prices of option and stock for 40 companies traded on the JSE. The aspects that form part of the theoretical framework relevant for this study are characteristics of financial time series data, historical volatility, time series stationarity, statistical arbitrage, pairs trading strategy, statistical cointegration relationship, distance trading method, stochastic spread and the combined forecast method.

2.2.1 Characteristics of financial time series data

Stylised facts is defined by Sewell (2011) as an economics term that refers to the empirical findings that are consistent and accepted as truth. Financial time series is characterised by stylised facts. The most common stylised facts are no presence of autocorrelation (also referred to as serial correlation), flat tails and gain/loss asymmetry. Serial correlation is tested using the following equations:

$$Q = n(n + 2) \sum_{k=1}^Z \frac{\hat{\rho}_k^2}{n-k} \quad (2.1)$$

where Q is the Ljung-Box test statistic, n denotes the sample size, $\hat{\rho}_k$ is the serial correlation at lag k and z is the number of lags being tested. The asymmetric distribution is tested using Jarque-Bera (JB) (1980) test presented in the following equation:

$$JB = n \left[\frac{s^2}{6} + \frac{(k-3)^2}{24} \right] \quad (2.2)$$

where

$$s = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}{\left[\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right]^{3/2}} \quad (2.3)$$

and

$$k = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\left[\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right]^2} \quad (2.4)$$

where s is sample skewness and k is kurtosis. According to Sewell (2011), the serial correlation of log returns/stock prices is generally not significant. Cont (2001) found the presence of serial correlation on a tick-by-tick for both foreign exchange and stock shares trading on the New York Stock Exchange (NYSE). This finding confirmed that this serial correlation feature cannot be isolated from the financial time series data. A normal/symmetrical distribution has a kurtosis value of 3 while a leptokurtic distribution has a kurtosis value of more than 3, the platykurtic distribution has a kurtosis value less than 3. A leptokurtic distribution is characterised by heavy tails. This means that the distribution places more observations on the tails of its support than a normal distribution. A platykurtic distribution is characterised by short tails (Tsay, 2005). Kat (2003) is of the view that most of the stocks or assets tend to have a negative skewness or excess kurtosis. The same view was also expressed by Ling (2006) and Ang and Chen (2002).

2.2.2 Historical Volatility

As discussed in Chapter 1, volatility is defined as a statistical measure of dispersion around a mean value; or as the changeability or randomness of the underlying asset (Schwert, 1990). This measure is usually viewed from three different perspectives, namely, historical (backward looking), implied (reflected in an option market price) or future/actual (forward looking). Volatility is further defined by Sewell (2011:7) as “the standard deviation of the change in

value of a financial instrument and is considered a proxy for risk.” Therefore, historical volatility basically involves computing the variance or standard deviation of stock returns in the usual way over some historical period. This then becomes the volatility forecast for all periods in the future. Historical volatility of the actively traded options on a certain stock can be calculated by analysts and statisticians and the volatility is used to calculate the price of a less actively traded option on the same stock. In historical volatility, only closing prices of stock option are required. In such a scenario, historical volatility is defined as the standard deviation of the daily price index return for a period of time.

Volatility can be modelled using the OLS, ARCH and GARCH specifications. Chatfield (2004) outlined that classical time series models such as ARCH, ARIMA, GARCH and many other extensions and variations are frequently used to explore the data and make predictions. The classical time series methods are normally used to describe the dynamic volatility on the financial time series but they do not explore the mean reversion of the paired stock prices. Therefore the use of cointegration through the OLS approach will assist in the identification the pair of stocks that move together in the long run and also with the characteristic mean reversion.

2.2.3 Stationarity

It is widely known in econometrics that most of the time series data, especially financial time series data, are nonstationary. Which means that those time series contain unit root. Using dynamic model for nonstationary time series could result in bias in the regression results as stated by Yule (1926). Yule (1926) refers to this unbiasedness of results as spurious regression. This spurious regression was further investigated by Granger and Newbold (1974). After the work by Granger and Newbold (1974), Engle and Granger (1987), Johansen and Juselius (1990) pioneered the technique that combines the nonstationary variables to produce equilibrium relationships by using error correction models.

The two commonly used types of stationarity are strict stationarity and weak stationarity. A strict stationarity is obtained if a joint and conditional distribution of a process are unchanged if displaced in time. If Y_t ; $t \in \mathbb{Z}$ is a time series, then a time series is a strict stationary if the distribution of $(Y_{t_1}, \dots, Y_{t_k})'$ and $(Y_{t_1+h}, \dots, Y_{t_k+h})'$ are equal for all k and all $t_1, \dots, k; h \in \mathbb{Z}$. In short, the narration is written as

$(Y_{t1}, \dots, Y_{tk})' \stackrel{\text{def}}{=} (Y_{t1+h}, \dots, Y_{tk+h})'$ where $\stackrel{\text{def}}{=}$ means “equal in distribution”.

A $Y_t; t \in \mathbb{Z}$ series is said to be weak stationary if

- $Var(Y_t) < \infty \quad \forall t \in \mathbb{Z}$ (2.5)

- $\mu_Y(t) = \mu \quad \forall t \in \mathbb{Z}$ (2.6)

- $Cov(Y_t, Y_s) = \gamma_Y(r, s) = \gamma_Y(r + t, s + t) \quad \forall r, s, t \in \mathbb{Z}$ (2.7)

If the mean (2.6), variance (2.5) and covariance (2.7) of a time series are independent of time, then the series is said to be weakly stationary. There are three types of nonstationarity processes (Brockwell and Davis, 2002). The first one is random walk with and without drift, the second one is a series with deterministic trends and the last one is a combination of stochastic and deterministic trends. The types of nonstationarity processes are discussed below.

- Random walk without drift

The random walk without drift is the type of stationarity process without the intercept. The idea behind this type of stationarity is that future values cannot be estimated. The process is given by:

$$Y_t = Y_{t-1} + \varepsilon_t \tag{2.8}$$

where ε_t is a white noise term, $E(Y_t) = Y_0$ and $Var(Y_t) = t\sigma^2$. The variance of the series depends on time (t) confirming that the process is nonstationary.

- Random walk with drift

The process for random walk with drift is similar to random walk without drift but the only difference is the inclusion of an intercept term which is referred to as drift. The drift makes the process to drift downwards or upwards depending on the sign of the drift. The process is given by the equation below:

$$Y_t = \alpha + Y_{t-1} + \varepsilon_t \tag{2.9}$$

where ε_t is a white noise term, α is a constant/drift, $E(Y_t) = Y_0 + t * \alpha$ and $Var(Y_t) = t\sigma^2$. The $E(Y_t)$ and $Var(Y_t)$ of the process depends on t . This means that the process is not stationary.

- Deterministic trend

A deterministic trend is a process that is a function of time (t) and it is predictable. A random walk with a drift is often confused for a deterministic trend. They both have a component of a drift and a white noise respectively. The difference is that the value at time of a random walk with a drift is regressed on the last period's value (Y_{t-1}), while the deterministic trend is regressed on a time trend (β_t). A nonstationary process with a deterministic trend has a mean that grows around a fixed trend, which is constant and independent of time. The deterministic trend process is expressed by the following equation:

$$Y_t = a + bt + \varepsilon_t \quad (2.10)$$

where ε_t is a white noise term, $E(Y_t) = a + bt$ and $Var(Y_t) = \sigma^2$. $E(Y_t)$ depends on time and therefore the process is nonstationary.

- Combination of stochastic and deterministic trends

In this random walk process, the AR(1) is combined with a deterministic trend. Even if the AR process does not contain unit root ($|\rho| < 1$), the series is still considered as a nonstationary process due to the presence of the deterministic trend. If the series contains a unit root, a series can be transformed into a stationary process by a procedure called differencing.

Since the nonstationary process is characterised by unit root, it has become a tradition to test for unit root. The classical examples of formal tests for the presence of unit root include the Phillips-Perron test (Phillips and Perron, 1988), Augmented Dickey-Fuller test (Dickey and Fuller, 1981), the Kwiatkowski-Phillips-Schmidt-Shin test (1992), the Elliot-Lothman-Stock point-optimal test (Elliot et al., 1996) and the Ng-Perron test (Ng and Perron, 1995). The study by Song and Chang (2003) revealed that overlooking stationarity resulted in model coefficients being inflated and model overestimation due to the presence of serial correlation. Wei (2006) stated that the problem of unit root can be solved by introducing the logarithms

and differencing the series. If a time series becomes stationary after d times of differencing, the process is referred to as an $I(d)$ series.

The unit root tests are generally examined for the null hypothesis of unit root. However, in order to model the dynamics of utmost financial time series, Augmented Dickey-Fuller (ADF) (1981) formulated a test that accommodates the more general autoregressive moving average form. In ADF test, it is assumed that error terms are homoscedastic and there is no presence of serial correlation. The Phillips-Perron (PP) (1988) test is a nonparametric test for stationarity. The null hypothesis tested under the PP test is that the series is non-stationary. The PP test statistic uses the same asymptotic distributions as the ADF test but the PP test has an advantage over the ADF test. The advantage of PP test is that it is more robust to error terms showing presence of heteroscedasticity and there is no need to specify the lag length when applying the test.

Ng-Perron (1995) came up with a test that allows the lag length to be more flexible so that it can change based on the sample size on condition that error terms follow the general autoregressive moving average process. The method proposed by Kwiatkowski-Phillips-Schmidt-Shin (KPSS) (1992) differs from the ADF and PP tests. The KPSS test assumes stationarity of the series in the null hypothesis. The study by Pfaff (2008) suggests that researchers should always follow the KPSS test since it puts the hypothesis of interest as the alternative as compared to other tests of stationarity. Once the variables have been tested for stationarity, a cointegration technique may be applied.

2.2.4 Statistical arbitrage

Statistical arbitrage is defined by Bondarenko (2003) as a zero-cost trading opportunity for which the average expected payoff is non-negative. The study by Hogan et al. (2004:526) defined statistical arbitrage as “a zero initial cost, self-financing trading strategy with cumulative discounted value.” The concept of statistical arbitrage is a generalization of the traditional “riskless” interpretation. Statistical arbitrage identifies mispricing based on deviations from common stochastic trends and relies on a predictive modelling framework that attempts to exploit consistent regularities in the movements of asset prices, unlike the traditional interpretation (Alsayed, 2014). The first working paper on statistical arbitrage was written by Gatev et al. (1999) and it was published seven years later in the *Review of Financial Studies*. However, statistical arbitrage had been known many years before 1999 on Wall Street.

Traditional approaches state that even though the price in the short run can be influenced by irrational traders (noise traders), the rational traders (arbitraders) will immediately exploit the attractive investment opportunity and implement the arbitrage strategy thereby correcting the mispricing (Sudak and Suslova, 2009). Statistical arbitrage is a pairs trading that has been used by hedge fund managers, professional traders and institutional speculators. This method exploits market inefficiencies by taking into account two highly correlated pairs of stocks (Perlin, 2009). Elliott et al. (2005) and Gatev et al. (2006) all consider statistical arbitrage as pairs trading.

According to Ehrman (2006), statistical arbitrage is based purely on historical statistical financial data that is utilized in very short term for numerous small positions and it is almost purely model and computer driven when any single trade has very little human analysis. Avellaneda and Lee (2010) indicated that pairs trading strategy is the “ancestor” of statistical arbitrage. Xie and Wu (2013) indicated that the main idea behind pairs trading is to take advantage of the temporary mispricing of two indexes that have a common historical movement.

Statistical arbitrage is defined by Caldeira and Moura (2013) as a technique that is based purely on the assumption that the historical observed patterns are going to be repeated in the future. Therefore, statistical arbitrage is a statistical approach designed to exploit equity market inefficiencies defined as the deviation from the long term equilibrium across the historically observed stock prices. Statistical arbitrage is therefore convincingly described by Lo (2010:260) as “a highly technical short-term mean-reversion strategy which involves large numbers of securities, short holding periods, substantial computational models, and trading.” Statistical arbitrage technique is strongly related to the cointegration technique since it depends on mean reversion (Meki, 2012).

The three broad components of statistical arbitrage are as follows:

- First component is to construct a statistical relationship between stock prices.
- Second component is to identify statistical arbitrage opportunities by forecasting the changes in combinations of stock prices.
- Third and last component is to implement the most appropriate trading strategy.

These three components that form base for statistical arbitrage methodology are discussed later in the methodology chapter. The main focus of this study is on the application of the statistical arbitrage technique in modelling option prices.

2.2.5 Pair trading strategy

The discovery of pairs trading was in the early 1980s by the quantitative analyst Tartaglia and the team of physicists, computer scientists and mathematicians who had no background in finance. Their discovery was to develop statistical rules that could help in performing arbitrage trades (Gatev et al., 2006). The pairs trading is often referred to as statistical arbitrage strategy or a market neutral trading strategy. The idea behind pairs trading strategy is to identify a pair of stocks prices that exhibit historical co-movement. A position is opened when there is a significant deviation from the historical relationship and close position when there is a convergence in the stocks (Gatev et al., 2006).

Pairs trading is defined by Ehrman (2006:167) as “a non-directional, relative-value investment strategy that seeks to identify two companies with similar characteristics (a pair) whose equity securities are currently trading at a price relationship that is outside their historical trading range. Pairs trading is a neutral long/short investment strategy. This means that investors do not have to forecast the future directions of the stock prices, but focus on the stocks that have a similar historic price patterns, without looking at the conditions of the market at any point in time (Georgios, 2013). This investment strategy entails buying the undervalued security while short-selling the overvalued security, thereby maintaining market neutrality.”

Vidyamurthy (2004) indicated that the conception of pairs trading is the simultaneous buying and selling two securities that are historically correlated. Pairs trading strategy requires buying the under-valued stock (long position) while short selling (short position) the over-valued stock, consequently keeping market neutrality. Vidyamurthy (2004:8) further stated that “pairs trading is said to be a market neutral strategy in its most primitive form which eliminates the effect of market movements when using just two securities, consisting of a long position in one security and a short position in the other, in a predetermined ratio.” According to Schmidt (2008), in order to make profit from this relative mispricing, a long position in the stock is opened when its value falls adequately below its long run equilibrium and is closed out once the value of the stock reverts to its expected value. In the same way, investors may earn profit

when a stock is trading sufficiently above its equilibrium value by withholding the stock until it reverts to its expected value.

The concept of pairs trading strategy is based on the relative mispricing of stock prices. Pairs trading is one of the commonly used strategy of statistical arbitrage. Pairs trading technique uses the relative values between the paired stocks to find statistical mispricing. The first step in implementing pairs trading strategy is to identify stocks that are moving together and having the same characteristics. This step is called “pairs formation phase.” The co-movement of stocks is identified using the correlation analysis and cointegration method. The second step spots the relative mispricing in the pairs (i.e. when the “price relationship is outside their historical trading range”). This second stage is called “divergence in the pair.” The last step is trading the pair that would mean-revert to its long run equilibrium and this stage is called “convergence in the pair.”

The deviation from the central point is indicated by spikes in the spread, suggesting trading opportunities. An entry point is calculated using the standard deviation. The convergence in the pairs is exploited by trading long in the overvalued and short in the undervalued stock and vice versa. This means that if the spread reaches a certain predetermined standard deviation, there is a profitable trading opportunity. Figure 2.1 by Yakop (2011) summarises the implementation of pairs trading strategy. The x-axis represents period of trading and y-axis represents the units of the historical relationship between paired stock prices.

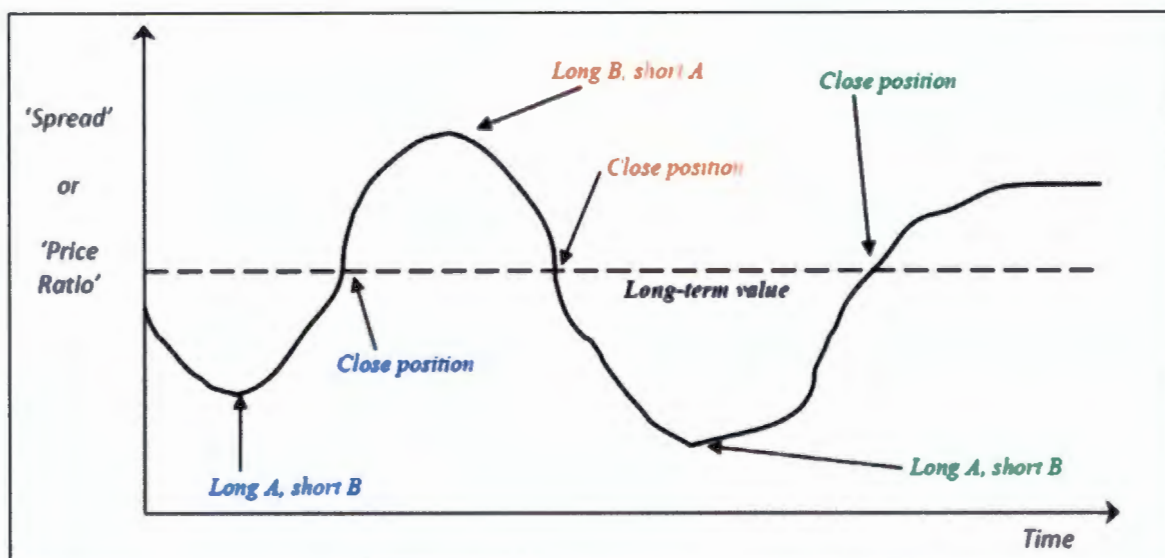


Figure 2.1: Graphical representation of pairs trading strategy (Yakop, 2011)

Pairs trading strategy that is based on the dynamics portfolio was developed by Vidyamurthy (2004). Elliott et al. (2005:271) are of the view that pairs trading is “an investment strategy that is built on utilizing relative mispricing between two securities whose historical prices have a pattern of co-movement. The argument is that if two security prices co-move, then they might be driven by the same underlying factors and, according to the law of one price, hence be priced in the same way.” Nicholas (2000:237) indicate that “the trading strategy of pairs trading is to hold an equal amount of long security positions and short security positions with the sole purpose of minimising the risk of the portfolio and at the same time making a profit.”

Pairs trading strategy involves modelling two related stocks whose relative pricing is away from a state of equilibrium. This technique is called pairs trading because pairs traders take two positions when they make the trade. The idea of pairs trading is to go long on the underperforming stock ($\mu - \Delta$) while simultaneously going short on the over performing stock ($\mu + \Delta$). The trade is closed once the position reverts back to its central point. Going long simply means buying the stock while going short means selling the over-priced stock. Hansell (1989) indicate that pairs trading is a relative value investment strategy where investors buy an “under-priced” stock while concurrently selling a similar “over-priced” stock. The stock prices are often selected on the basis of having a similar historical movement. The pairs trading strategy is often viewed as a market neutral strategy.

Constructing the trading threshold is vital in the performance of pairs trading strategy. Trading thresholds are set to guide an investor on when to enter trade and when to exit. Let P_t^A and P_t^B be the prices of stock A and B respectively at a given time t . The trading rule can be constructed such that to open trade an investor will go short if $P_t^A - \hat{\beta}P_t^B - \hat{\mu} \geq q$ simultaneously go long if $P_t^A - \hat{\beta}P_t^B - \hat{\mu} \leq -q$. The trade is closed when the spread reaches central point (μ). The standard trading rule, as stated by Gatev et al. (2006), is the two standard deviation threshold. According to standard trading rules, if the price ratio is above (below) two standard deviations, there is an opportunity to make profit when trading short (long) the numerator and long (short) the denominator.

The rationality behind pairs trading is that an investor looks for pairs of stocks that move together and share similar characteristics. Once there is a divergence in the paired stocks from the mean level, then there is high probability that they will revert back to their mean level again. In the econometrics literature, the mean level is usually referred to as the spread (Rudy et al., 2010). Basically, the concept of pairs trading is the simultaneous buying and selling of two historically correlated and or cointegrated stock indexes. Generally speaking, these strategies exploit the relative mispricing of the two assets by taking a long position in the over-priced asset and a short position in the under-priced one, while maintaining market neutrality by setting long/short positions. There are four main methods in implementing pairs trading strategy, namely:

- Cointegration method
- Distance trading method
- Stochastic spread method
- Combined forecast method

The four main methods are discussed in some detail in the following subsections.

2.2.6 Cointegration

According to Alexander (1999), in the application of statistical arbitrage technique, correlation analysis and cointegration technique are closely related, but they address different notions. Alexander and Dimitriu (2015) indicated that the application of cointegration technique to financial econometrics has increased over the years and the technique has been proven to be significantly effective. Highly correlated pairs of stock does not imply that stock prices are cointegrated in the long run. Correlation simply depicts the co-movements in stock prices. On

the other hand, cointegration models the long run relationship in stock prices even when the stock prices are not strongly correlated. Huck and Afabuwo (2015:602) stated that “cointegration is based on the idea of mean reversion and it attempts to find the long term relationship between asset prices.” According to Chiu and Wong (2012:516), “cointegration is a useful econometric tool for identifying assets which share a common equilibrium. Cointegrated pairs trading is a trading strategy which attempts to take a profit when cointegrated assets depart from their equilibrium.”

The concept of cointegration was pioneered by Granger (1981) and Engle and Granger (1987) and it is widely used in stock markets. The cointegration applicability to stock markets was initiated by Lucas (1997) and Alexander (1999). Cointegration technique can model both the long and the short run relationship between variables (Alexander, 1999). This is done by tracking historical data to check whether the series would move together and then assume that the data would continue as such in the future. Cointegration technique has two implications when modelling statistical arbitrage. The first implication is to determine the long run relationship between a set of stock prices which could be considered as the statistical equivalence of a “fair price.” The deviation from the long run relationship could be regarded as “statistical mispricing.” The second implication is the mean reverting component in the time series. In short, cointegrated pairs trading is a strategy which attempts to profit from statistical mispricing.

Engle and Granger cointegration techniques are used to test the statistical relationship between two time series data that are integrated to same order d , $I(d)$, to produce a single time series which is integrated of the same order $d - b$, where $b > 0$. The Engle and Granger (1987) test incorporates the long run equilibrium theory into the model so that any disequilibrium can be corrected. The Engle and Granger (1987) test is a residual based tests used to examine the presence of unit roots on residuals of single regression equation models in order to test the null hypothesis of the absence of cointegration. The test by Engle and Granger (1987) examines the presence of cointegration using a two-step approach as discussed below:

- The first step is to estimate the cointegration relationship using the following ordinary least squares equation:

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t \quad (2.11)$$

and test for cointegration by applying the ADF test to the residuals from the estimated regression. The y-intercept is denoted by β_0 , β_1 is the slope and ε_t is the error term. The residuals are estimated as follows:

$$\Delta \hat{\varepsilon}_t = \rho \hat{\varepsilon}_{t-1} + \sum_{i=1}^{n-1} \omega_i \Delta \hat{\varepsilon}_{t-i} + \rho_t \quad (2.12)$$

where $\rho_t \sim i.i.d N(0, \sigma^2)$ and $\hat{\varepsilon}_t$ are estimated from the above equation.

- Once the results reveal that variables are cointegrated, the next step is to construct the error correction model (ECM) and use the lagged residuals from the cointegrating relationship in step 1 to correct the disequilibrium.

Since the current study applied the Engle and Granger cointegration technique and as such the bilateral ECM was modelled. Another method of testing for cointegration was developed by Johansen (1988). The Johansen (1988) developed the maximum likelihood test, which is an alternative method of establishing the long run cointegration between variables. The method developed by Johansen is based on the vector autoregressive process. The Johansen cointegration (1988) test has the advantage over the Engle and Granger two step method since it can determine the existence of more than one cointegration relationship between the variables. Alexander (1999:2043) postulated that “the Johansen cointegration uses the likelihood estimator to determine a linear combination that is most stationary given a set of variables to be tested for cointegration.” According to Granger (1983:6), “the relevance of cointegration lies in the fact that the parameter β_1 can be optimised from historical prices using a cointegration regression. Using the cointegration approach, the “fair price” relationship is modelled by regressing the historical stock price of asset X on the historical stock price of asset Y.”

Hendry (1986) made a significant contribution in the direction of testing Granger causality. The Granger technique test whether adding lags of one variable has a predictive power for another variable. Granger (1987) indicated that the presence of cointegration suggests causality running from at least one direction among the variables.

2.2.7 Distance trading method

The distance trading method was first introduced in the study of markets and commodities by Gatev et al. (1999). According to Gatev et al. (1999), the distance trading method is established in discussions with investors passively using a pairs trading approach. The initial step in the distance approach is to express the daily returns for each stock price as a ratio of stock price (D_{it}) to its value on the preceding day (D_{it-1}). The ratio is calculated using equation (2.13) below.

$$A_{it} = \log \left(\frac{D_{it}}{D_{it-1}} \right) \quad (2.13)$$

The returns are then normalised by subtracting the sample mean of the ratio stock prices and dividing this by their sample standard deviation. The process is given by the following computation:

$$A_{it}^Z = \frac{A_{it} - E(A_{it})}{s_i^A} \quad (2.14)$$

where A_{it} denotes normalised prices of asset i at time t , $E(A_{it})$ is the expectation of A_{it} and s is the standard deviation. The pairs are then selected by computing the differences in the sum of squares between the normalized stock prices at a particular period of time. The selected pairs are then ranked from the highest to the lowest order using differences in their sum of squares. The differences in the sum of squares is calculated using the following equation:

$$D = \sum_{t=1}^N (A_{it}^X - A_{jt}^Y)^2 \quad (2.15)$$

where D denotes the cumulative differences in the sum of squares between the normalised stock prices and A^Z denotes the normalised stock prices. The pairs with the lowest sums are of greatest potential in a pairs trading strategy. Do et al. (2006) pointed out that the disadvantage of distance trading approach is nonparametric in nature and it does not allow the researchers to estimate errors in the model. Nevertheless, the distance approach also relies on the properties of the cointegration trading approach.

If the potential pairs are identified, then the spread between the paired stocks is computed. The spread (S_t) is computed using the following equation:

$$S_t = S_t^X - S_t^Y \quad (2.16)$$

The trade is opened when S_t exceeds some predetermined value q . In precise terms, an investor would go short in X and simultaneously go long in Y when $S_t > q$. An investor would go short in Y and simultaneously go long in X when $S_t < -q$. The trade is then closed when S_t converges to the predetermined closing position.

2.2.8 Stochastic spread method

The common approach using the stochastic approach was outlined by Elliott et al. (2005). In a stochastic spread method, the spread between two stocks is modelled as a stochastic variable with mean reverting properties. Elliott et al. (2005) used the following equation to model the spread between two stocks:

$$y_k = x_k + D\omega_k \quad (2.17)$$

where x_k represents value of a variable at a given time $t_k = k\tau$ for $k = 0, 1, 2, \dots$, ω_k is independent and identically distributed (i.i.d.). Gaussian $\sim N(0,1)$ and constant measure of error is denoted by D and it is greater than 0. It is assumed that x_k follows a process given by:

$$x_{k+1} - x_k = (a - bx_k)\tau + \sigma\sqrt{\tau}\varepsilon_{k+1} \quad (2.18)$$

where $\sigma \geq 0, b > 0, a \in \mathcal{R}$ and ε_k is i.i.d. Gaussian $\sim N(0,1)$ and independent of ω_k from equation (2.17). The $x_{k+1} - x_k$ in equation (2.18) will mean revert around $\mu = a/b$ with “power” b . If $Y_k = \sigma(y_0, y_1, y_2, \dots, y_k)$ represents observed information from $y_0, y_1, y_2, \dots, y_k$, the conditional expectation in equation (2.19) represents the estimate of the hidden process of equation (2.13) through the observed process of equation (2.17):

$$\hat{x}_k = E(x_k | y_k) \quad (2.19)$$

Equation (2.19) can be rewritten as follows:

$$x_{k+1} = A + Bx_k + C\varepsilon_{k+1} \quad (2.20)$$

where A, B and C are the coefficients of the model. $A = a\tau \geq 0$, $0 < B = 1 - b\tau < 1$ and $C = \sigma\sqrt{\tau}$. The x_k can be regarded as $\cong X(k\tau)$, where $X(t)|t \geq 0$ satisfies the following stochastic differential equation:

$$dX(t) = (a - bX(t))dt + \sigma dW(t) \quad (2.21)$$

The Ornstein-Uhlenbeck process is used to approximate equation (2.21) in order to estimate A, B and C to obtain equation (2.20). An investor can go short when $y_k \geq q\hat{x}_{k|k-1}$ and simultaneously go long when $y_k \leq -q\hat{x}_{k|k-1}$. The threshold value for when trades are open is denoted by q . Do et al. (2006) opined that the disadvantage of the model is that it restricts the long run relationship between the stock prices.

2.2.9 The combined forecast method

The combined forecast method was developed by Huck (2009, 2010). This author used the combination of two methods, namely, the multi-criteria decision making methods (MCDM) and neural networks methods to test pairs trading strategy by using Standard and Poor's (S&P) 100 stocks. The framework of the combined forecast method is based on three stages, namely, forecasting, ranking and trading. In the first step, a universe of n stocks are considered so that $\frac{n(n-1)}{2}$ combination of pairs can be constructed. Huck (2009, 2010) used the Elman neural networks to generate the one week ahead return forecast for each stock i with condition to the past returns of stock i and j with $i, j \in (1, \dots, n)$. Huck (2009, 2010) used the MCDM in the second stage called "ELECTRE III." In the second stage, a set of alternatives were ranked to a set of criteria. The performance (x_{ij}) of each stock (i) relative to criterion j can be estimated using the following equation:

$$x_{ij} = \hat{x}_{|X_{i,t}, X_{j,t}}^{i,i+1} - \hat{x}_{|X_{i,t}, X_{j,t}}^{j,j+1} \quad (2.22)$$

The x_{ij} values are collected in an antisymmetric $n \times n$ matrix. The undervalued stocks are placed on top of the ranking while the overvalued stocks are placed at the bottom of the rank. In the last step, the stocks ranked at the top are bought and the bottom ranked stocks are sold short. The positions are closed after trading for one week then a new ranking is created and the process is repeated every week.

The combined forecast method is developed without reference to any equilibrium model. Huck (2009, 2010) state that the method offers much more trading opportunities and could detect the “birth” of the divergence that other methods cannot achieve. The shortfall of the combined forecast method is that the pairs do not share any equilibrium relationship.

2.3 EMPIRICAL LITERATURE

The academic literature on pairs trading strategies is very limited and scarce in both qualitative and quantitative studies to date. Only a few studies have been conducted using pairs trading methodology. Gatev et al. (1996) are the first authors to document pairs trading investment strategy that existed for over twenty five years on Wall Street. Then other authors have started documenting pairs trading statistical arbitrage investment strategy. The published research on pairs trading is still very limited.

Gatev et al. (2006) conducted a study to test a pairs trading strategy using the distance method for data ranging from 1962 to 2002. Their study formulated the pairs using a twelve month data period. The pairs were formed by minimising the squared deviations between the normalised price series of the two stocks. Gatev et al. (2006) used the general rule of thumb (two standard deviation rule) in deciding whether to enter trade or not. The selected pairs were subsequently opened when their spread reached two standard deviations from their historical mean and the pairs were closed when the spread returned to its historical mean. The pairs trading strategy used by Gatev et al. (2006) created an annual returns of over 10%. The study by Gatev et al. (2006) was reproduced by Andrade et al. (2005) using the security exchange for Taiwanese market. The study used daily time series data for the period 05 January 1994 to 29 August 2002. This Andrade et al. (2005) study sought to determine if their study results would be similar to those reported by Gatev et al. (2006). The pairs trading strategy by Andrade et al. (2005) produced similar results of an average annual return of 10%. Andrade et al. (2005) further stated that profit may be generated from shocks in pairs trading in stocks in the Taiwanese market.

The study by Schroder and Smith (2001) implemented pairs trading strategy on the 200 most liquid stocks in Europe. The study differs from Gatev et al. (2006) as Schroder and Smith (2001) formulated the pairs using the two step approach, namely, the ADF test and the beta test. The trading rule applied by Schroder and Smith (2001) is simple two standard deviation rule just like in Gatev et al. (2006). The difference between the study by Schroder and Smith (2001) and Gatev et al. (2006) is that Schroder and Smith (2001) added an extra restriction that they open trade when the spread of a pair crosses the two standard deviation band for the second time. This means that they opened trade when the spread is on its way back from the second two standard deviation band.

The study by Perlin (2007) tested the pairs trading strategy of the 57 most liquid stocks from the Brazilian market between the year 2000 and 2006. The study used the same technique of selecting pairs as Gatev et al. (2006). Ruiter (2011) tested statistical arbitrage pairs trading strategy to the 100 most liquid stocks exchange of Japan, Korea and Taiwan. The study by Ruiter (2011) employed daily data ranging from 2000 to 2009. The pairs formed were based on the log share price ratio being cointegrated over a period of two years and they were traded for a period of six months. Similar trading used by Schroder and Smith (2001) was also used by Ruiter (2011). An extra restriction was also added: trade was opened when the spread of a pair crosses the two standard deviation band for the second time. Out of all the sectors, only 5.71% of the pairs that crossed the two standard deviation barrier for the first time were restricted to never open. Interestingly Ruiter (2011) was very careful in his risk minimising approach and only considered trading when the spread crosses the deviation for the second time. Even though the proportion of the pairs that did not cross the two standard deviation for the second time was very small, it was very interesting to note that Ruiter (2011) concentrated only upon those that crossed for the second time to guard against all the risks.

The study by Hong and Susmel (2003) implemented pairs trading strategies on 64 Asian shares listed in their local markets and in the United States (US) as American depository receipts. This study applied cointegration method to detect the pairs. The study revealed that an investor waits for a year before closing trade-in pairs trading. Likewise, Puspaningrum et al. (2009) used cointegration technique to select and to implement pairs trading strategy. Their study estimated the number of trades over a specified trading horizon. Moreover, Lin et al. (2006) tested pairs trading strategy on the cointegrated series of the daily closing price data for two Australian Stock Exchange quoted bank shares (Australia New Zealand bank and Adelaide bank) from

January 2001 to August 2002. Lin et al. (2006) found that cointegration pairs trading approach yielded minimum profit. The focus of the study by Lin et al. (2006) was not to maximise profit but to ensure that the pairs trading strategy yielded minimum profit. Stock trading is a risky business and therefore it is very important to choose as a trader whether to expose yourself to high risk or low risk. High profit is normally associated with high risks and the loss could be very substantial. Lin et al. (2006) were very cautious in how they traded for less risk and less profit but high sustainability over time.

Ackert and Tian (1999) examined the efficiency of the S&P 500 index options market in the U.S using daily closing prices from 01 January 1986 to 31 December 1996. This study revealed that the option market inefficiencies are reduced over time, but that the introduction of depository receipts, restrictions on short sales and transaction costs incurred affect the profitability of index option arbitrage negatively. On the statistical side, the pricing reliance again reverts to the degrees of correlation between the share or index prices. Where the share or index prices are highly correlated, the relationship between their volatility levels can be calculated according to Ammann and Herriger (2002). In the light of that fact, the study by Ammann and Herriger (2002) revealed that the price relationship that is created through the correlation of the underlying share prices can be extended to the volatilities and therefore to the behaviour of the prices of the options on the underlying shares.

Ammann and Herriger (2002) examined the statistical arbitrage relationship that exists between United States' eleven share indexes. Their approach was to remove stationarity using standard stationarity tests. This removed one index from the eleven tested. The next test was to identify indices that were correlated to a degree greater than 0.95, and this removed a further 5 indexes. Of the five remaining indices, the three most liquid were used in their analysis. The findings of their study indicated their trading strategy to be effective, with profitable trades being in the 95% levels for all indexes traded.

Muslumov (2009) tested the pairs trading strategy in Istanbul stock exchange using the distance approach. The author normalised the stock prices before implementing pairs trading strategy. Nath (2003) employed the distance approach in the liquid secondary US stock market from January 1994 to December 2000. In this study, trade was opened when the distance between the stocks widened to reach or cross 15% trigger level. Trade was closed when the spread of the distance reached or crossed its median or when the last day of trading period is reached. A

similar approach was used by Bianchi et al. (2009) on the returns on daily commodity. Their study found that pairs trading in similar commodity futures produced statistically excess returns. Monitoring of the commodities is high amongst the restricted sector trade. More or excess returns would be realized because there would not be any destruction as would be expected if one traded in the unrestricted sectors.

The study by Bolgun et al. (2009) used daily data for the period 2002 to 2008 on the Istanbul stock market to perform pairs trading technique. The study implemented the distance approach and did not group stock prices into sectors. This study found that pairs trading technique produced average returns of 3.36% for the sample period. The results further revealed that pairs trading strategy produces more returns with less volatility as compared to the market portfolio. Cummins (2010) only used the distance approach to examine pairs trading strategy of the nine industry sectors classified by Bloomberg. The study by Cummins (2010) found that utilities was the worst performing sector. The study further found similar results when comparing the industry-restricted pairs and unrestricted pairs.

Miao (2014) used high frequency stock data from US stock market between 2012 and 2013. The author used two step approach to identify potential pairs. The pairs were preselected using correlation coefficients. These preselected pairs were tested for cointegration to identify the best pairs. The cointegrated pairs were subsequently traded when the deviations from the estimated relationships arose. Similarly, Bowen, Hutchinson and O'Sullivan (2010) modelled pairs trading using high frequency data derived from the United Kingdom (UK) stock market for the year 2007. The authors followed the approach used by Gatev et al. (2006) and their study found moderate excess returns. The study by Kim (2011) also modelled pairs trading strategy using high frequency data from the Korea stock market. The author selected pairs belonging to the same industry sector. From the study by Kim (2011), it was found that there are positive returns in all market conditions with better performance in bear market conditions. The study also found that performance of pairs trading strategy is more related to the timing in market entry. The similarity in the three above studies is the application of pairs trading strategy using high frequency data.

A seminal paper by Engle and Granger (1987) introduced the cointegration quantitative technique used to find the long term relations between stock prices. Cointegration is a technique used to identify potentially related stock prices. Johansen (1988) developed another approach

which could be alternatively used when dealing with more than two stock prices at the same time. Alexander (2001) indicated that the Engle and Granger approach is the more preferred method in financial applications than other variants due to its simplicity and lower variance.

Vidyamurthy (2004) formulated pairs trading models using Engle and Granger two step cointegration approach. In this study, Vidyamurthy (2004) demonstrated how to generate trading signals when relative price deviates sufficiently from its long run equilibrium. This is just a practitioner's guide to pairs trading and there are no empirical results shown or discussed in this guide.

The study by Caldeira and Moura (2013) used cointegration technique as a technique for selecting stocks for pair trading strategy. The paper used data from the Sao Paulo stock exchange for the period January 2005 to December 2010. Out of 1225 possible pairs, 90 pairs were cointegrated. The pairs trading implemented in the study showed relatively low levels of volatility. There was a high level of stability in the sector trading, suggesting that low levels of volatility increase investor confidence in the said sector.

The study by Koronidis (2013) tested the pairs trading technique in forex markets. The study used Engle and Granger cointegration methodology to test whether or not the EUR/USD and GBO/USD currency pairs were cointegrated. The results in this case revealed that the pairs were not cointegrated at a 5% level of confidence. Therefore, pairs trading was not possible since the pair was not cointegrated.

Similarly, Nath and Verma (2003) used both Engle and Granger cointegration and multivariate cointegration to examine the long run relationship between stock prices for India, Singapore and Taiwan. The study found that there is no cointegration between the stock indexes. In the same breadth, Hoel (2013) used Engle and Granger cointegration approach on the Norwegian stock market to back-test the performance of pairs trading technique for a sample period ranging from 2003 to 2013. The study revealed that the implementation of back-test resulted in large losses. The Engle and Granger cointegration technique was also used by Galenko et al. (2012) to obtain the mean-reverting signals, and evaluated a pre-assumed convergence trading strategy on a handful of exchange-traded funds tracking equity indices. Galenko et al. (2012) found that profit is possible when implementing cointegration pairs trading strategy.

Pairs trading strategy was tested by Schmidt (2008) on the Australian stock market. Pairs of stocks were identified using Johansen cointegration technique and modelled the residuals using the Vector Error Correction Model (VECM). The results revealed a high rate of zero crossing and larger deviations around the mean. This means that a profitable pairs trading was possible in the Schmidt (2008) model. Similarly, the study by Chang (2002) employed the multivariate trace statistic, an approach that had been used by Harris-Inder and the Johansen cointegration technique to model the relationship between Shanghai and Shenzhen stock exchange. Their study used daily closing prices for the period January 1995 to December 1999. The study by Chang (2002) found that the two stock markets are not cointegrated. Kasa (1992) also employed the Johansen cointegration technique to investigate the co-movement between the US, Japan, England, Germany and Canada stock markets between 1974 and 1990 and it was found that there was indeed a common stochastic trend throughout the sample period.

Chiu and Wong (2012) studied the optimal dynamic of the pairs trading strategy of cointegrated assets by using the mean-variance portfolio selection criterion. The study suggested that arbitrageurs do not trade when the stock prices are not cointegrated. The paper by Ungever (2015) employed daily time series data of 10 agricultural future markets ranging from 2004 to 2015 to examine pairs trading strategy by using the Johansen cointegration approach. Out of 45 agricultural futures market pairs, 7 were found to be cointegrated. The results of pairs trading strategy therefore found that only 2 pairs showed any trading signal. The trading rule used in the paper by Ungever (2015) is 1.5 standard deviations.

Haque and Haque (2014) implemented a pairs trading strategy to develop a financially profitable pairs trading model in the Dhaka stock exchange. These authors used daily closing prices of the 20 most liquid stocks listed in Dhaka stock exchange. The variables used were classified according to sectors. The pairs were identified using Johansen cointegration method. The VECM series was used as a guide to implement pairs trading strategy. Three pairs of stocks were found to have a long run equilibrium relationship. The trading rule used by Haque and Haque (2014) is that an investor will go long on the underperforming stock when the value of the residual is significantly below zero and go short on the outperforming stock when the value of the residual is significantly higher than zero. The trade is closed when the value of the residual is close to or reaches zero.

The study by Walls (1994) examined the relationship between natural gas spot prices at various production fields, pipeline hubs and city markets in US natural gas industry using Johansen cointegration test. According to the results of this study, there is a strong correlation between the natural gas spot markets. Out of nineteen market pairs, thirteen satisfied the rule for perfect market integration. Wong et al. (2005) studied the relationship between Indian stock market and the stock markets in the developing countries (US, UK and Japan) using Johansen multivariate cointegration test. Their study revealed that the Indian market is closely cointegrated with US, UK and Japan stock markets.

Bogomolov (2010) used three estimation methods, namely distance method, cointegration method and stochastic method to compare the profitability of pairs trading strategy in the Australian stock market. All the methods revealed statistically significant monthly excess returns on committed capital before transaction costs. The pairs trading strategy was also examined by Baronyan et al. (2010) using the distance, cointegration and the stochastic spread method. Their study found that pairs trading strategy performs better under severe market conditions.

Similarly, the study by Xu (2013) implemented pairs trading strategy only using the distance approach and the cointegration approach. The study selected pairs within the same industry and within different industries. It was found from the study that unrestricted pairs performed better than the restricted pairs when using the distance approach and it was also found that restricted pairs performed better for service, financial and retail sector when using the cointegration approach. The study by Xu (2013) also found that cointegration approach yielded higher excess returns as compared to the distance approach at the cost of high volatility. Similarly, Caldas et al. (2014) applied pairs trading strategy using cointegration and distance approach with data from Brazil and European stock market ranging from 1996 to 2012. The results from the study by Caldas et al. (2014) are similar to those from Xu (2013).

Liew and Wu (2013) used the distance approach, cointegration approach and copula approach to compare results of highly correlated stock pairs. The study used the sample period from 01 December 2009 to 30 November 2012. The results revealed that a copula method yielded highest number of trading signals as compared to the other two methods. A similar study by Augustine (2014) implemented pairs trading strategy using copula method, distance method

and the cointegration method with five different pairs of stocks listed on JSE. The results of the study are similar to those arrived at by Liew and Wu (2013).

The source of profit from pairs trading strategy was analysed by Engelberg et al. (2009:27). The study revealed that “the pairs trading profitability is positively linked to the way information spreads across the stocks in the pair and in the frictions that suppress this information flow.” Jacobs and Weber (2011) used the US stocks and 8 major stock markets to model returns in different information settings. Their study used a sample from 1960 to 2008 and it found that higher returns are generated when pairs are opening on high distraction days than when pairs are opened on low distraction days. In the study by Kanamura et al. (2010), the pairs trading strategy was applied to energy futures market for the period 2000 to 2008. Kanamura et al. (2010) used a mean reverting process of the futures price spread and their study reported that unchanging profit can be made from pairs trading strategy but the cross commodities’ profit may not be improved.

The study by Dunis et al. (2010) applied a pair trading strategy to the constituent shares of the Eurostoxx 50 index. The study implemented a basic long-short trading strategy to pairs of shares belonging to the same sector. The trading rule used by Dunis et al. (2010) was the 2 standard deviations away from the long term mean. A different approach was taken by Papadakis and Wysocki (2007) to model the impact of accounting information events such as earnings announcements, on the profitability of pairs trading strategy proposed by Gatev et al. (2006). The study found that trades are mostly triggered around accounting information events for the portfolio of US stock pairs between 1981 and 2006.

Mori and Ziobrowski (2011) did a comparative study for the period 1987 to 2008 to compare the pairs trading performance in the US REIT market with the US general stock market. The study revealed that the REIT market produced more profit opportunities between 1993 and 2000 but the profits decreased after 2000.

The study by Huang et al. (1996) used daily time series to investigate the relationship between the oil futures and stock returns using the vector autoregressive method. It was reported here that individual stock returns can be affected by oil future returns. A similar study by Buyuksahin et al. (2008) examined the relationship between commodity investment and commodity price using dynamic correlation and recursive cointegration for the period 1992 to

2008 and ultimately found that there is no change in the returns and there is no increase in the cointegration between the securities.

Lee et al. (1985) examined the relationship between the S&P500 and the Commodity Future Index. The authors found no relationship between the two series and they concluded that arbitrage opportunities are not possible since the pairs are not related. The relationship between Exchange Traded Funds (ETFs) and equities was investigated by Avellaneda and Lee (2008) using two different categories of statistical arbitrage trading technique, namely, principal component analysis (PCA) and sector ETFs. Lee et al. (1985) revealed that the two strategies yielded the best results on either 15 EFTs or a 15 PCA.

2.4 CHAPTER SUMMARY

Relevant recent and extant literature was reviewed and comparisons/differences drawn from the various studies. This literature reports on different techniques in modelling the pairs of stock. Most of the studies on pairs trading strategy were conducted on an international perspective and of those reviewed, none was conducted from a South African perspective. Therefore the current study tries to fill the gap in the implementation of pairs trading in a South African context. Several studies used cointegration technique in selecting pairs to be traded. The current study preselected the stock pairs using Pearson correlation coefficient analysis prior to cointegration analysis. This strategy was also applied by studies such as Miao (2014), Gatev et al. (2006), Herlemont (2004), Liew and Wu (2013) and Schmidt (2008). The above mentioned studies are mainly non-South African studies but they provide systematic modelling for application of the same processes to establish the reality within South Africa.

The study by Xu (2013) and Caldas et al. (2014) found that cointegration technique performed better than the distance approach while Liew and Wu (2013) and Augustine (2014) found that copula approach performed better than distance approach and cointegration approach. Distance approach was used by Cummins (2010) to examine pairs trading strategy across nine industry sectors and found the same results for industry-restricted and unrestricted pairs. Similarly, Xu (2013) also found that unrestricted pairs performed better than the restricted pairs when using the distance approach and that restricted pairs performed better for service, financial and retail sectors when using the cointegration approach. The current study implements pairs trading using four different sectors: financial sector, industrial sector, mining sector and retail sector. Most of the preceding studies found that there is a long run relationship between their

respective stock prices while the study by Koronidis (2013), Nath and Verma (2003) and Chang (2002) found out that their paired stocks were not cointegrated. Therefore pairs trading was not possible. It was revealed from the literature that investors would profit from the cointegrated pairs.

According to the literature, most of the authors used daily financial time series data to model statistical arbitrage trading techniques. Therefore, the current study takes the similar route of using daily time series data. Gatev et al. (2006), Schroder and Smith (2001) and Ruiter (2011) used the general rule of thumb (two standard deviation rule) as compared to Ungever (2015) in deciding whether or not to enter trade. Ruiter (2011) also factored an additional restriction of opening trade when the spread crossed the two standard deviation band for the second time. Do et al. (2006) contend that the stochastic spread model restricts the long run relationship between the stock prices and therefore is disadvantageous. Therefore cointegration approach enabled the study to model the long run relationship between the paired stock prices.

In the current study, the selected pairs are subsequently opened when their spread reaches two standard deviations from their historical mean and the pairs are closed when the spread returns to its historical mean. This means that the study applies the general rule of thumb as operationalized by Gatev et al. (2006) who are the most reviewed authors in pairs trading strategy.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 INTRODUCTION

This chapter discusses the research methodology employed in the study. The methodology follows a research onion model proposed by Saunders et al. (2012) to address the objectives specified in the first chapter. The methodology of this study is divided into broad components and pairs trading method is implemented thereafter. The two broad components are simple rolling regression and Engle and Granger cointegration. The methodology adopted ranges from univariate to multivariate techniques. This methodology was implemented to model historical volatility of index options of four different sectors to achieve the following objectives that were set to:

- identify pairs of highly correlated index options,
- determine whether or not there is a long run relationship (cointegration) between the selected stock indexes,
- determine the direction(s) of causality between the paired stock indexes,
- identify which pairs of stock are to be traded.
- use the findings of the study in formulating suggestions for investor decision purposes.

The chapter is organised as follows: Section 3.2 outlines the ethical considerations. Section 3.3 outlines the research process, Section 3.4 deals with identifying number of pairs, Section 3.5 focuses on data analysis framework, Section 3.6 presents the univariate procedures that consists of preliminary analysis, Section 3.7 presents the selection of indexes, Section 3.8 is the test for stationarity. Section 3.9 establishes the relationship between index pairs, followed by boundaries for regression coefficients in Section 3.10, then a section on pairs trading using cointegration techniques which include the error correction model is presented in Section 3.11, Section 3.12 is the error correction model, Section 3.13 outlines the model diagnostic, Section 3.14 outlines the causality test, Section 3.15 looks at the implementation of the statistical arbitrage trading strategy in South Africa.

3.2 ETHICAL CONSIDERATIONS

There is no ethical consideration related to environment, animals or human subjects in this study. The study only uses secondary data. However, the researcher remains truthful to the

recommendations of the institutional manual of postgraduate studies for the handling and manipulation of data in the execution of the study. Permission to conduct the study was sought from the university by submitting an application form to the research ethics committee through the supervisor after the defense and approval of the proposal.

3.3 RESEARCH PROCESS

The general research process is described by Saunders et al. (2012) using the onion metaphor. The author indicated that every research process should start by explaining the philosophical perspective which the research follows. This section discusses the research philosophy by specifically paying attention to research approach, research strategy, research choice, time horizon and data collection methods. The research onion displayed in Figure 3.1 gives direction for a research process. This onion clearly illustrates pertinent issues when undertaking a research.

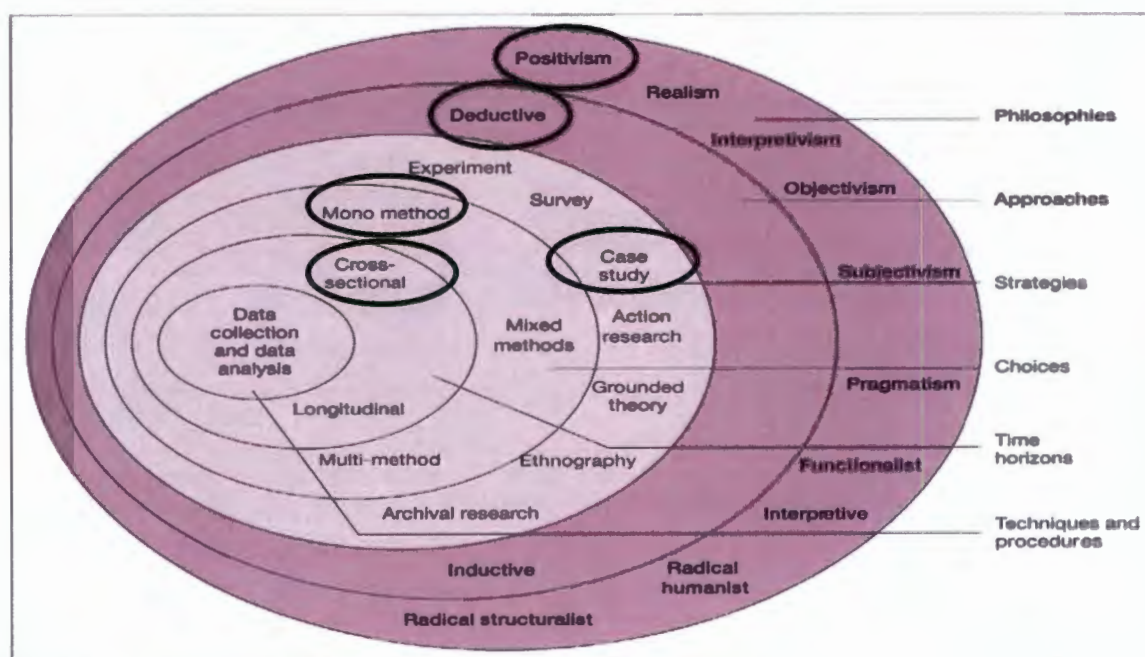


Figure 3.1: Research onion by Saunders et al. (2012)

3.3.1 Philosophy

According to Saunders et al. (2012), research philosophy is significant as it is what a researcher does when they are carrying out research in a topic: they are increasing their understanding in a certain field effectively. Philosophy layer outlines a number of philosophies available to the researcher. Saunders et al. (2012) further expounded on the three philosophical thoughts

through which researchers lay claim to knowledge such as ontology (what is knowledge), epistemology (how we know it) and axiology (what values go into it). A research question is the foundation of the epistemological, ontological and axiological orientation of the researcher (Rohlfers and Zhang, 2016).

According to Collis and Hussey (2009), epistemology refers to what constitutes acceptable knowledge. There are two elements in epistemology, namely, positivism and interpretivism. Walsham (1995b) explains positivism as the scientific knowledge consisting of facts. As Karnevio (2007) asserts, an interpretive paradigm forms a class of the sociology of regulation. The purpose of this paradigm is to seek an understanding of the world from an individual's viewpoint. Karnevio (2007) asserts that such a paradigm is nomothetic, positivistic, voluntaristic and ideographic using subjective first-hand knowledge. The advantage, according to Burrell and Morgan (2005), is that a researcher tries observing continuous developments with the aim of understanding individual conduct and the spirituality of the world. Interpretivism is described by Gephart (1999) as directed on meanings and understandings of the social interactions between humans.

Hwang (1996) views positivism as a broad variety of theories and practices. Therefore, the most appropriate philosophical stance for this study is positivistic epistemology. The study seeks answers to theory-driven questions or objectives (Creswell, 2013; and Creswell and Tashakkori, 2007) as outlined in Chapter 1 hence the applicability of this philosophy. According to Hussey and Hussey (1997) interpretivism and positivism are two extremities but they lie on similar continuum (class of epistemology) due to the fact that one tends to be applicable when qualitative data and a small sample is used and the other when quantitative data and a large sample are in question. Rubin and Babbie (2016) view a paradigm as a frame of reference which shapes people observations and understandings.

3.3.2 Research approach

The second layer of the research onion is the research approach. According to Saunders et al. (2012) and Saunders and Bezzina (2015), research approach constitutes the deductive and inductive approaches to research. Saunders et al. (2009) indicated that a deductive approach starts by formulating hypotheses which are later tested while in the inductive approach theories are derived from data analysis. This study obtained and used time series data which requires the quantitative methods for analyses. Therefore, the study followed a deductive approach since

it sought to answer research questions stated in Chapter one (Rocco et al., 2003). The study applies statistical arbitrage strategy to test the efficiency of stock markets in pricing relative risk in highly correlated stocks option traded on the JSE.

3.3.3 Research strategy

The next layer of Saunders' et al. (2012) research onion is about the research strategy adopted. The authors suggested seven strategies that a researcher can adopt depending on the nature of their studies. Case study is defined by Robson (2002:178) as "a strategy that involves the empirical investigation of a specific contemporary phenomenon." The research strategy used in this study was a case study since the study is conducted in a South African context using four sectors trading on JSE as experimental units. The sectors referred to are financial sector, industrial sector, mining sector and retail sector.

Robson (2002), and Hussey and Hussey (1997) defined a case study as a strategy for conducting a research that involves empirically investigating certain contemporary phenomenon in the context of its real life and using multiple sources of evidence. A phenomenon of interest whose prominence is critical could extensively be examined through case study. This study uses secondary data on a particular variable for South Africa as a case study as a research strategy. The fact that the study conducted is an empirical investigation into the South African top 40 index option also qualified a case study as a research strategy. The study further developed conclusions based on the data analysis (Yin, 2003a) making the proposed strategy a relevant candidate for the current research.

The fifth layer of the onion before the core is the time horizon. Time horizon has two elements which are cross sectional and longitudinal studies. Cross sectional studies are described by Saunders et al. (2009) as portraying an image of a selected phenomenon at a specific time period, while longitudinal studies represent events over a long period of time. The study is undertaken to answer the predetermined research questions and address a specific problem at a particular time, therefore cross-sectional strategy was considered.

3.3.4 Choice of research

Generally, there are two main research approaches, namely, qualitative and quantitative research. According to Creswell (2012), qualitative research investigates that which consists of getting information (data) from the participants by means of a questionnaire. Quantitative

research is defined by Maree (2012) as a structured way of data collection in numbers from a certain group extracted from a population and analysing that data in order to get the findings that represent the studied population. Based on the two definitions, this study followed a quantitative approach since daily time series data obtained from the JSE was used.

Since this study utilised quantitative data and approaches, the choice of research is narrowed to mono-method. Van der Merwe (1996) defined a quantitative method as an approach aimed at testing theories, determining facts, determining associations among variables and prediction of outcomes. This research approach uses approaches from the natural sciences designed to ensure objectivity, generalizability and reliability (Weinreich, 2010). The current study used statistical methods to achieve predetermined objectives regarding the relations between specific variables, as a result, quantitative methods are suited for the enquiry.

Daily time series data used in this study covered the period 04 January 2010 to 31 December 2015 and consist of 1500 observations with one year approximated to have 250 trading days. The stock markets do not operate on weekends and on public holidays, therefore the sample period excluded weekends and public holidays. The choice of the time period is based on the availability of data from the source, JSE. The study used the stock market data comprising closing prices of option stock for 40 companies traded on the JSE. As highlighted in previous sections, the data comprise four main sectors, namely: financial, industrial, mining and retail sectors. These sectors were chosen based on their relative size as well as their importance to the South African economy. The data forms a good experimental unit for the application of statistical arbitrage. This data is also relevant and suited for the set objectives and study significance. Researchers from other countries such as Schroder and Smith (2001), Gatev et al. (2006), Perlin (2007) and Ruiter (2011) have also used similar data on studies involving the application of statistical arbitrage application. Statistical software package used was E-Views. These software provide results relevant to achieve the study objectives. The stock market data together with the stock code provided by JSE are presented in Table 3.1.

Table 3.1: Stock market data

<i>Financial Sector</i>	<i>Industrial Sector</i>	<i>Mining Sector</i>	<i>Retail Sector</i>
Investec Limited (INL)	African Rainbow Minerals Gold (ARI)	Anglo American (AGL)	The Foschini Group (TFG)
Barclays Africa Group Ltd (BGA)	Nampak (NMK)	Anglo American Platinum (AMS)	Aspen Pharmacare Holdings (APN)
Bidvest Group (BVT)	Sasol (SOL)	AngloGold (ANG)	British American Tobacco PLC (BTI)
Firststrand Limited (FSR)	Vodacom Group (VOD)	Arcelor Mittal South Africa Ltd (ACL)	The Spar Group (TSG)
Standard Bank Group (SBK)	Telkom (TKG)	BHP Billiton (BIL)	Woolworths Holdings (WHL)
Sanlam (SLM)	Sappi (SAP)	Gold Fields Limited (GFI)	Truworths International (TRU)
RMB Holdings (RMH)	Netcare Limited (NTC)	Impala Platinum Hlds (IMP)	Tiger Brands (TBS)
Old Mutual (OML)	MTN Group (MTN)	Pretoria Portland Cement (PPC)	Shoprite (SHP)
Nedbank Group (NED)	Mondi Ltd (MND)	Kumba Iron Ore (KIO)	SABMiller (SAB)
Discovery Ltd (DSY)			Pick'n Pay Stores (PIK)
Imperial Holdings (IPL)			
Reinet Investments (REI)			

3.4 IDENTIFYING NUMBER OF PAIRS

In mathematics, the binomial coefficient is referred to as any positive integer that occurs as a coefficient in the binomial theorem (Hilton and Pedersen, 1991). The total number of pairs can be described by the binomial coefficient given by the following equation:

$$P\binom{n}{r} = \frac{n!}{r!(n-r)!} \quad (3.0)$$

where n denotes sample size and r is the group size. Therefore, using the formula in equation (3.0) there are 66 $\binom{12}{2}$ possible different pairs that can be created under financial sector, 36 possible different pairs under industrial sector, 36 possible different pairs under mining sector and 45 possible different pairs under retail sector.

3.5 DATA ANALYSIS FRAMEWORK

The analysis of data for the study followed the framework summarised in Table 3.2. Similar methods of data analysis were applied to individual sectors.

Table 3.2: Analysis framework

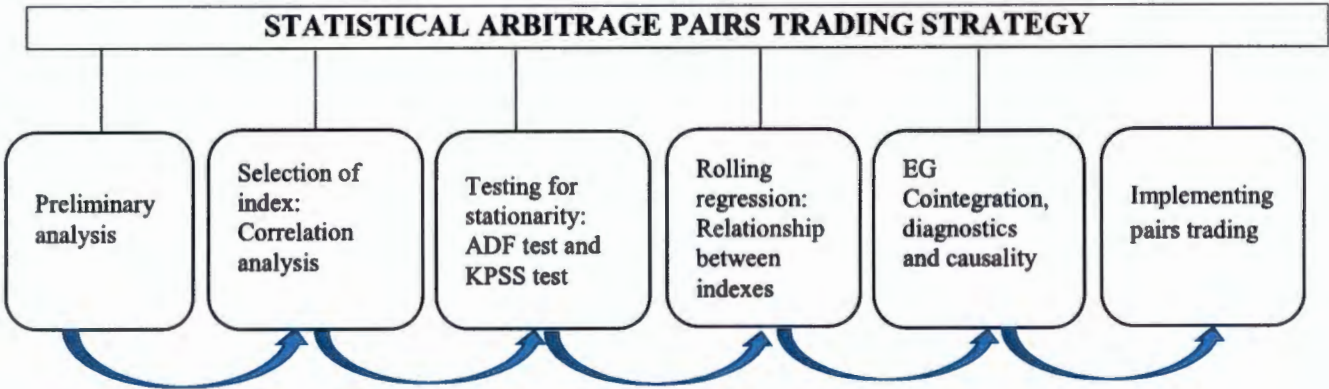


Table 3.2 summarises the analysis framework of the study. The framework is further thoroughly discussed in the following sections.

3.6 PRELIMINARY ANALYSIS

Preliminary data analysis gives a clear understanding of the characteristics of the financial time series data used in the study. Financial time series is characterised by stylised facts. The most common stylised facts are no presence of autocorrelation, flat tails and gain/loss asymmetry (Sewell, 2011). Most of the financial data are highly volatile and they have leptokurtosis effect. This means that the series has features of flat tails to the normal distribution (Yonis, 2011). The normal distribution function has a kurtosis value of 3 and leptokurtosis distribution has a kurtosis value of more than 3 while platykurtic has a kurtosis value of less than 3. The kurtosis and Pearson’s measure of skewness coefficient are given by equation (3.58) and (3.59) respectively. The kurtosis value of financial time series ranges between 5 and 1000 (Cont, 2001). The asymmetry of the distribution is measured using skewness while flat tails are measured using kurtosis. The stylised facts of this study are derived from the descriptive statistics. All this descriptive statistics are discussed in detail in Section 3.13 for convenience purpose.

3.7 SELECTION OF INDEXES

The criteria for selection of indexes is based on the assumption that all the selected indexes should be highly correlated. One other assumption is that the index pairs are linearly related to one another. The associated tests for these assumptions are discussed in subsequent sections.

3.7.1 Correlation analysis

The correlation analysis examine the strength of the relationship between two highly correlated indexes that have price trends. Pair traders seek to trade instruments that tend to move together. In selecting each of the paired index, all of the indexes should belong to the same sector. In this study, correlation analysis is viewed as a pre-selection criteria prior to cointegration analysis. The criteria of selecting indexes belonging to the same sector is motivated by studies by Gatev et al. (2006), Herlemont (2004) and Schmidt (2008). The authors stated that using pairs that are highly correlated over time lead to successful implementation of pairs trading strategy. The problem that could arise from highly correlated stocks is a spurious relation, stock could appear to be related while they are not. To address this shortfall, unit root tests were used to overcome the problem of spurious correlation.

The use of correlation criterion is motivated by the assumption that index pair with high correlation reveal a strong linear relationship between each other. Profit is created by any changes in correlation that may be followed by mean reversion to the trend of stock pairs (Miao, 2014). Therefore, before establishing any relationship, pairwise correlations were tested. The study used correlation coefficient of 0.90 as a cut-off point³.

Considering stock X and Y, a correlation coefficient between stock X and Y is a statistic that only provide a measure of how the two stocks are related in a sample but also its properties, which closely related them to a straight line regression.

The population Pearson's correlation coefficient of stock X and Y is defined as:

$$\rho_{XY} = \frac{cov(X,Y)}{[V(X)V(Y)]^{1/2}} \quad (3.1)$$

where

³ The choice for using 0.90 as a minimum value for correlation coefficient is arbitrary, it is guided by the fact that coefficient should be as high as possible and some indexes should be left out in further analysis.

$$cov(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [X - E(X)][Y - E(Y)] f(X, Y) dXdY \quad (3.2)$$

and

$$V(X) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [X - E(X)]^2 f(X, Y) dXdY \quad (3.3)$$

where

$$E(X) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} X f(X, Y) dXdY \quad (3.4)$$

$V(Y)$ and $E(Y)$ are similarly defined in terms of Y . The quantity ρ_{XY} is a measure of the linear associated between the random variable X and Y . If a sample of size n , $(X_1, Y_1), (X_1, Y_2), \dots, (X_n, Y_n)$, is available from the joint distribution. The sample Pearson's correlation coefficient, r , of stock X and Y is obtained by the following equation:

$$r_{XY} = \frac{\sum_i^N (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{[\sum_i^N (X_i - \bar{X})^2 \sum_i^N (Y_i - \bar{Y})^2]}} \quad (3.5)$$

where $\bar{X}$ and $\bar{Y}$ are mean prices of stock X and Y respectively. The formula for $\bar{X}$ and $\bar{Y}$ are given by:

$$\bar{X} = \sum_i^N \frac{X_i}{N} \quad (3.6)$$

$$\bar{Y} = \sum_i^N \frac{Y_i}{N} \quad (3.7)$$

where N denoted a stock trading data range and r denotes a correlation coefficient. The value of r ranges between -1 and +1, where:

- Perfect negative correlation (-1) exists if two stocks move in opposite directions (i.e. stock X moves up while stock Y moves down);
- No correlation (0) exists if the pairs of stocks are completely random (i.e. stock X and Y go up and down randomly) and
- Perfect positive correlation (+1) exists if the paired stocks move in perfect unison (i.e. stock X and Y move up and down at the same time).

The strong positive r is the more positive is the association of stock X and Y. This means that stock X and Y are highly matched and they should be selected for further analysis. A high risk is involved when trading pairs of stock since the probability of a break in the correlation may occur while trading is high. When a break occurs while a pair is “open” this would most likely lead to a loss. Therefore, the risk is minimised by selecting the pairs of stocks that belong to the same sector (Chan, 2009 and Dunis et al., 2010). To avoid spurious regression, the data was transformed and tested for stationarity or lack thereof. The next subsection discusses the data transformation.

3.7.2 Data transformation

Osborne (2010:1) indicated that “data transformations are commonly-used tools that can serve many functions in quantitative analysis of data, including improving normality of a distribution and equalizing variance to meet assumptions and improve effect sizes. Data transformation constitute important aspects of data cleaning and preparation for your statistical analyses.” The frequently used method of data transformation is the conversion of data into natural logarithms. The natural logarithm ($\ln(x)$) is the logarithm having base e . If $e^x = y$, then natural log function is defined as:

$$\ln(x) = \log_e(x) \quad (3.8)$$

where $e = 2.7182818 \dots$

The function e^x is defined as the exponential function and the inverse of the exponential function is the natural logarithm for $x > 0$. The following are the rules and properties of natural logarithm:

- $\ln(xy) = \ln(x) + \ln(y)$
- $\ln(x/y) = \ln(x) - \ln(y)$
- $\ln(x^y) = y \ln(x)$
- $f(x) = \ln(x) \Rightarrow f'(x) = \frac{1}{x}$
- $\int \ln(x) dx = x(\ln(x) - 1) + c$
- $\ln(x)$ is undefined when $x \leq 0$
- $\ln(0)$ is undefined and $\ln(1) = 0$

The study takes into account the rules and properties of natural logarithm when transforming the data. The data used in this study was transformed into natural logarithms in order to stabilise sample variability and prepare the data for statistical analyses. Using natural logarithms allows the differences to be expressed as returns. The discussion in the next step relates to the two proposed stationarity test statistics.

3.8 TESTS FOR STATIONARITY

The tests for unit roots are important in determining whether time series is stationary or not because a regressor that is nonstationary invalidates various empirical results. The presence of a stochastic trend is determined by modeling the presence of unit roots in time series data. If the series is nonstationary and its first difference is stationary, then the series contains a unit root. To avoid spurious correlations and regression, the study tested each of the selected index time series for stationarity. Unit root test was only applied to the preselected pairs. This was done not only to allow the use of statistical arbitrage method, but also to establish the appropriateness of cointegration which also forms part of this study. The proposed Augmented Dickey–Fuller (ADF) by (Dickey and Fuller, 1979, 1981) and Kwiatkowski, Phillips, Schmidt and Shin (KPSS) by (Kwiatkowski et al., 1992) tests were applied to the series to determine their order of integration. The following properties characterise a stationary time series for all t and $t - s$:

$$E(X_t) = E(X_{t-s}) = \mu \quad (3.9)$$

$$E[(X_t - \mu)^2] = E[(X_{t-s} - \mu)^2] = \sigma_X^2 \quad (3.10)$$

$$E[(X_t - \mu)(X_{t-s} - \mu)] = E[(X_{t-j} - \mu)(X_{t-j-s} - \mu)] = \gamma_s \quad (3.11)$$

where μ , σ_X^2 and γ_s are constants. Equations (3.9), (3.10) and (3.11) therefore state that a stationary series has a constant mean, variance and autocovariance (Enders, 2010). Equation (3.9) is the most important property in terms of pairs trading since if the spread between two stock prices are found to have a constant mean, any deviations from this value can be traded. A framework for testing unit root is represented by the following equation:

$$X_t = \phi X_{t-1} + \varepsilon_t \quad (3.12)$$

then X_{t-1} is subtracted from both sides to obtain:

$$X_t - X_{t-1} = \varphi X_{t-1} - X_{t-1} + \varepsilon_t \quad (3.13)$$

$$\Delta X_t = (\varphi - 1)X_{t-1} + \varepsilon_t \quad (3.14)$$

$$\Delta X_t = \delta X_{t-1} + \varepsilon_t \quad (3.15)$$

where ε_t represents a serially uncorrected white noise error term with a mean of zero and a constant variance, Δ denotes first differenced operator. If $\delta = 0$, then φ will be 1, implying that equation (3.12) is a random walk without drift model, which means the process is nonstationary. If the process is nonstationary, the researcher is faced with the problem known as the unit root. However, if $\delta < 0$ then $\varphi < 1$, then the X_t series is stationary. The stationarity of the series is important because correlation could continue in nonstationary time series even if the sample is very large and may have effect in what is called spurious or nonsense regression (Ssekuma, 2011). The problem of unit root can be solved, or stationarity can be achieved, by differencing the dataset (Wei, 2006). The ADF test and KPSS test are discussed in subsequent sections. Before testing the relationship between the two indexes, the classical regression model requires that all the variables used in the study be stationary, this implies that their means and variances remains constant over time (Ssekuma, 2011).

3.8.1 Augmented Dickey-Fuller (ADF) Test

The ADF tests the null hypothesis of unit root in a series, suggesting that the series is not stationary. The ADF test refers to the t-statistics of δ_i coefficient on the following regression equation:

$$\Delta Z_t = \alpha + \beta t + \gamma Z_{t-1} + \sum_{i=1}^p \delta_i \Delta Z_{t-i} + \mu_t \quad (3.16)$$

where α and β are the coefficient on a trend, ΔZ_{t-i} represent the first difference with p lags and μ_t is the variable that adjusts the errors of autocorrelation. The coefficients γ and δ_i are computed using the following equations respectively:

$$\gamma = -(1 - \sum_{i=1}^p \alpha_i) \quad (3.17)$$

and

$$\delta_i = -\sum_{j=i}^p \alpha_j \quad (3.18)$$

If $\alpha = 0$ and $\beta = 0$, equation (3.16) models a random walk. If $\alpha \neq 0$ and $\beta \neq 0$, equation (3.16) models a random walk with a drift. The Ordinary Least Square regression (OLS) is used to estimate all the parameters in equation (3.16). The null and alternative hypotheses for ADF test are stated as follows:

$$H_0: \delta = 0$$

versus

$$H_1: \delta < 0$$

The ADF test is a lower-tailed test, so if the p-value of the test statistic is less than the 1%, 5% or 10% level of significance, then the null hypothesis of unit root is rejected and it can be concluded is that the variable of the series does not have a unit root and it is nonstationary. In order to determine the optimal length (p), the Akaike Information Criterion (AIC) and the Schwarz Bayesian Criterion (SBC) are used. Non-rejection of the null hypothesis indicates that the series is nonstationary, whereas the rejection of the null hypothesis implies that the series is stationary.

3.8.2 Kwiatkowski, Phillips, Schmidt and Shin (KPSS) Test

Kwiatkowski et al. (1992) proposed a test of the null hypothesis of stationarity against the alternative of a unit root. KPSS test is implemented as a complementary procedure to the ADF test. The KPSS test is a Lagrange Multiplier test for stationarity. The KPSS test is derived from the following model:

$$y_t = \varphi t + \rho_t + \varepsilon_t \tag{3.19}$$

$$\rho_t = \rho_{t-1} + \mu_t \tag{3.20}$$

where deterministic trend is denoted by t with coefficient y and ρ . The error process, ε_t , is a $I(0)$ process from equation (3.19) and μ_t is the error process of equation (3.20). The null and alternative hypotheses for KPSS test is as follows:

$$H_0: \sigma^2 = 0$$

versus

$$H_1: \sigma^2 > 0$$

The test statistic is given by:

$$LM = \frac{\sum_{i=1}^T S_t^2}{\hat{\sigma}_t^2} \quad (3.21)$$

where $\hat{\sigma}_t^2$ represents the long run error variance estimated from the regression of y_t on a constant and time t and S_t is the partial sum of the residuals $\hat{\varepsilon}_t$ from the same regression. S_t is estimated by the following equation:

$$S_t = \sum_{i=1}^t \varepsilon_i, t = 1, 2, 3, \dots, N \quad (3.22)$$

and

$$S^2 = N^{-1} \sum_{i=1}^N \varepsilon_i^2 + 2N^{-1} \sum_{s=1}^l \left(1 - \frac{s}{(l+1)}\right) \sum_{t=s+1}^N \varepsilon_t \varepsilon_{t-s} \quad (3.23)$$

The ε_t 's represents the residuals obtained by regressing the series being tested on a constant, N represent the number of observations and l is the lag length. S^2 is non-negative. The critical values are given in KPSS (1992). If the p-value of the KPSS test statistic is greater than the 5% level of significant, the null hypothesis is rejected in favour of the alternative. Meki (2012:43) stated that “the reason stationary processes are so attractive in asset investment is that they exhibit the mean reversion property. This property simply says that when stock prices increase (decrease) to their maximum (minimum) level then the market forces them down (up) towards an equilibrium constant.”

3.9 RELATIONSHIP BETWEEN INDEX PAIRS

Regression analysis aims to establish the relationship between two or more variables. Relationship between any pair of variables is denoted by x and y (Wegner, 2016). For each and every pair of the selected indexes, the daily return of one index are regressed onto the daily returns of the other using the OLS technique. The OLS technique may be used to cater for time-varying dependencies between the pairs of selected indexes. In order to allow for time-variant regression coefficient, Tesfatsion and Kalaba (1989) introduced the flexible least square as a generalization of the standard linear regression model. In order to solve the problem of time variation, the current study models historical volatility using the flexible OLS approach which is also called the rolling OLS approach. The rolling OLS regression is used to establish the

relationship between each pair of highly correlated indexes. The regression equation is given as follows:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{\mu}_i \quad (3.24)$$

where Y_i is daily returns of index Y at time t , X_i is daily returns of index X at time t , $\hat{\beta}_1$ and $\hat{\beta}_2$ are regression coefficients and $\hat{\mu}_i$ is the disturbance term. Where $\hat{\beta}_1$ and $\hat{\beta}_2$ is estimated by equation (3.22) and (3.23).

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x} \quad (3.25)$$

$$\hat{\beta}_2 = \frac{\sum_{i=1}^n y_i x_i - \frac{(\sum_{i=1}^n y_i)(\sum_{i=1}^n x_i)}{n}}{\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}} \quad (3.26)$$

where $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ and $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$. For this study, the sample used in every case is half a year of daily returns. According to Ammann and Herriger (2002:7), “using six month sample is long enough not to be seasonally bias but short enough to remain somewhat flexible to structural changes.” One year is approximated as having 250 trading days, therefore half a year will comprises of 125 trading days. Therefore, this creates a rolling regression of 125 days, where the oldest data points are dropped every time a new ones are added (Ammann and Herriger, 2002). The t-tests and associated p-values are used to test the significance of the regression coefficients. The hypothesis to test the significance of $\hat{\beta}_2$ is stated as:

$$H_0: \hat{\beta}_2 = 0$$

versus

$$H_1: \hat{\beta}_2 \neq 0$$

The test statistic is given by

$$t_{(\frac{\alpha}{2}, n-2)} = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)} \quad (3.27)$$

Model significance consist of comparing the p-values to the 5% significance level and rejecting the null hypothesis when the p-value is less than the significance level. This means that the null hypothesis is rejected in favour of the alternative hypothesis if $|t| > t_{(\frac{\alpha}{2}, n-2)}$.

3.10 BOUNDARIES FOR REGRESSION COEFFICIENTS

The slope of the coefficient is calculated and used to determine the relationship between the selected pairs using the respective 125 trading days as a sample (as indicated in equation (3.24)). The point estimators are subject to estimating error because coefficients might vary over time. The upper and lower boundary should be considered to render the estimated slope of the coefficients more robust as a forecasting tool.

The boundaries of the estimated slope coefficient is calculated as follows:

$$\hat{\beta}_i \pm t_{(\frac{\alpha}{2}, n-2)} * SE(\beta_i) \quad (3.28)$$

where $\hat{\beta}_i$ is the estimated slope coefficient and SE is the standard error. The boundaries are calculated using 95% confidence interval. Confidence interval gives information about the unknown parameter. If the confidence interval includes zero, one can conclude that the population is not significantly different from zero at a given confidence interval (Knezevic, 2008). After relationship has been established using simple OLS, the next step tests whether the paired stock have a long run relationship or not. The long run relationship is tested using the Engle and Granger (1987) cointegration technique.

3.11 PAIRS TRADING USING COINTEGRATION TECHNIQUE

The cointegration technique is used to check if the preselected stock pairs using the correlation coefficients share any long term equilibrium relationship. The concept of cointegration as an innovative mathematical model in econometrics and finance was developed by Engle and Granger (1987). This idea assumes that the two stock prices follow a common stochastic trend. The spread between these variables may be weakly stationary. The definition of cointegration is as follows:

- it assumes that X_t and Y_t are two time series that were nonstationary and
- there exists a parameter γ such that the following equation:

$$Z_t = \ln Y_t - \gamma \ln X_t \quad (3.29)$$

was a stationary process, the natural log of X_t and Y_t would be cointegrated. Cointegration provides a sound methodology for modelling both the long term equilibrium and short term dynamic trends of the time series. Let P_t^A and P_t^B be the prices of stock A and B respectively at a given time t . The log price of stock A is regressed against the log price of stock B. If the $\ln P_t^A$ and $\ln P_t^B$ were nonstationary, parameter γ existed such that the following equation was a stationary process:

$$\ln(P_t^A) - \gamma \ln(P_t^B) = \mu + \varepsilon_t \quad (3.30)$$

where μ is cointegration mean, ε_t is a stationary mean-reverting process. ε_t is referred to as a regression residual or a cointegration residual and γ represent the cointegration coefficient. Equation (3.29) represent a model of a cointegrated pair of stock A and B.

The cointegration process determined the cointegration coefficient γ and the mean μ of the cointegration model is determined by the long term equilibrium relationship between stock A and B. Therefore, a quantity of Q of a profit or loss in pairs trading could be estimated as follows:

$$Q = (P_t^A - P_{t+1}^A) - \gamma(P_t^B - P_{t+1}^B) \quad (3.31)$$

$$Q = (P_t^A - \gamma P_t^B) - (P_{t+1}^A - \gamma P_{t+1}^B) \quad (3.32)$$

$$Q = (\mu + \varepsilon_t) - (\mu + \varepsilon_{t+1}) \quad (3.33)$$

$$Q = \varepsilon_t - \varepsilon_{t+1} \quad (3.34)$$

The possible trading results in terms of profit or loss for each stock pairs trading process are created in equation (3.31). If $\varepsilon_t - \varepsilon_{t+1} > 0$, the pairs trading made a positive profit. If $\varepsilon_t - \varepsilon_{t+1} = 0$, the pairs trading was a breakeven. If $\varepsilon_t - \varepsilon_{t+1} < 0$, then pairs trading produces a negative returns.

3.11.1 Verification of cointegration

The study applies the Engle and Granger (1987) two step approach to test for cointegration in each paired stocks to find out whether there is a long run relationship between them. The Engle and Granger two step approach is commonly used when testing the relationship between two stock prices. Cointegration enables the researcher to combine the two stocks so that the combined indexes are stationary. Pairs trading technique is used to exploit short term deviations from a long run equilibrium relationship between the two stock prices. Asteriou and Hall (2011) indicated that if variables are cointegrated, their long run relationship is not spurious.

The first step in Engle and Granger (1987) method is to set up a cointegration regression between stock A and B as stated in equation (3.29), then estimate the regression parameters μ and γ using the OLS method. Afterward, the regression residual ε_t is tested for stationary or not. If the regression residual ε_t is stationary, then the two stock prices (P_t^A, P_t^B) are said to be cointegrated. The Engle and Granger (1987) cointegration is done by employing the ADF test. The long run relationship between two stock prices (P_t^A, P_t^B) is estimated using the following OLS equation:

$$A_t = \beta_0 + \beta_1 B_t + \varepsilon_t \quad (3.35)$$

where β_0 is a constant and β_1 is the coefficient of the cointegration. The ADF stationarity test is used on the regression residual ε_t to determine whether it has a unit root. The ADF test is given by equation (3.16). The residual series is estimated by the following:

$$\hat{\varepsilon}_t = A_t - \beta_0 + \beta_1 B_t \quad (3.36)$$

The number of lag order p in equation (3.16) is estimated using minimum information criterion. The number of lag order p is determined using the information criteria for selecting the appropriate lag order, such as the Akaike information criterion (AIC) (Akaike, 1992), Hannan-Quinn criterion (HQC) (Hannan and Quinn, 1979), Schwartz information criterion (SIC) (Schwarz, 1978), Bayesian information criterion (BIC) and Final Prediction Error (FPE) (Akaike, 1969) (Akaike, 1979; Liew, 2004). The technique for estimating the number of lag order p is to minimize one of the information criterions. The method for selecting the optimal length p is to minimise one of the following information criterion:

$$AIC = \ln(\hat{\sigma}_p^2) + \frac{2p}{T} \quad (3.37)$$

AIC was derived by Akaike (1973) as an approximately unbiased estimator of the Kullback-Leibler information.

$$HQC = \ln(\hat{\sigma}_p^2) + \frac{2p \ln[(T)]}{T} \quad (3.38)$$

HQ was derived by Hannan and Quinn (1979) based on the approach adopted by Shibata for developing the Shibata's information criterion.

$$SIC = \ln(\hat{\sigma}_p^2) + \frac{p \ln(T)}{T} \quad (3.39)$$

Schwarz (1978) used a Bayesian perspective to derive SIC. The Schwarz information criteria as SIC is referred to by some researchers, others call it the BIC, whereas some combine the terms and refer it as the Schwarz Bayesian criterion (SBC).

$$BIC = (T - p) \ln \left(\frac{T \hat{\sigma}_p^2}{T - p} \right) + T[1 + \ln(\sqrt{2\pi})] + p \ln \left[\frac{\sum_{t=1}^T (\Delta Z_t)^2 - T \hat{\sigma}_p^2}{p} \right] \quad (3.40)$$

and

$$FPE = \hat{\sigma}_p^2 (T - p)^{-1} (T + p) \quad (3.41)$$

Akaike's Final Prediction Error (FPE) criterion provides a measure of model quality by simulating the situation where the model is tested on a different data set. According to Akaike's theory, the most accurate model has the smallest FPE.

T in equations (3.37), (3.38), (3.39), (3.40) and (3.41) is the sample size and $\hat{\sigma}_p^2$ represent the error variance. The formula for error variance is given by:

$$\hat{\sigma}_p^2 = \frac{\sum_{t=p}^T \hat{\mu}_t^2}{T - p - 1} \quad (3.42)$$

where μ_t is the error term presented in equation (3.11). Therefore, using one of the information criterion, called $C_{T(p)}$ in equations (3.37), (3.38), (3.39), (3.40) or (3.41) the optimal lag order p could be estimated using the following equation:

$$\hat{p} = \operatorname{argmin}_{p \leq p_{\max}} (C_{T(p)}) \quad (3.43)$$

the $p_{\max}$ represent the maximum lag order. The null and the alternative hypotheses for the existence of unit root in variable ΔX_t are as follows:

$H_0: \sigma_2 = 0$ (Not cointegrated)

versus

$H_1: \sigma_2 < 0$ (Cointegrated)

A statistical value for the ADF test is given by

$$ADF = \frac{\hat{\gamma}}{SE(\hat{\gamma})} \quad (3.44)$$

where $\hat{\gamma}$ and $SE(\hat{\gamma})$ are the cointegrating coefficient and standard errors of the OLS estimate respectively. The standard error $SE(\hat{\gamma})$ is computed using the following equation:

$$SE(\hat{\gamma}) = \sqrt{\frac{\sum_{i=1}^n (\Delta Z_{ti} - \Delta \hat{Z}_t)^2}{(n-2) \sum_{i=1}^n (Z_{(t-1)i} - \hat{Z}_{(t-1)i})^2}} \quad (3.45)$$

The ADF test result in equation (3.44) will be compared with the critical value of the ADF test. If the test statistic is less than the critical value or the p-value of the test statistic is less the 1%, 5% or 10% level of significance then the null hypothesis will be rejected. This means that there is no existence of unit root and the regression residual in equation (3.30) is stationary. Therefore, the log of stock prices (P_t^A, P_t^B) are cointegrated.

3.12 ERROR CORRECTION MODEL (ECM)

According to Granger representation, when two time series are cointegrated, a vector autoregressive (VAR) model on the differences is mis-specified (Granger, 1986). This mis-specification can be resolved by including the previous disequilibrium term as an exploratory variable. By doing so, the model will become well specified. The ECM is a dynamic model

that can reconcile the long run equilibrium with disequilibrium behaviour in the short run, which allows testing for short term or dynamic causality. After obtaining the disequilibrium term from equation (3.35), the ECM is applied to two cointegrated series $\ln P_t^A$ and $\ln P_t^B$:

$$\Delta A_t = \alpha_1 + \gamma_1 Z_{t-1} + \varepsilon_{1t} \quad (3.46)$$

$$\Delta B_t = \alpha_2 + \gamma_2 Z_{t-1} + \varepsilon_{2t} \quad (3.47)$$

where Z is the disequilibrium term given in equation (3.44), $Z_t = \beta_0 + \beta_1 B_t + \varepsilon_t$. Constant is represented by α_1 and α_2 respectively while γ_1 and γ_2 denote the speed of adjustment coefficients of the equilibrium. The OLS Regression applied on equation (3.46) and (3.47) by running the regression of the first difference of A_t and B_t on the first lagged disequilibrium term.

Lags are added to the ECM to ensure that residuals show no sign of autocorrelation. Alexander (2008) stated that γ_1 and γ_2 also indicates how fast the short term deviation will move back to the long term equilibrium following an exogenous shock. Enders (1995:328) indicated that “the larger γ_1 or γ_2 is, the quicker the response of the dependent variable to the deviation from long run equilibrium of the previous period’s disequilibrium. When the coefficients are small, the speed of adjustment is slow, in which the dependent variable does not respond to the previous period’s disequilibrium.”

3.13 MODEL DIAGNOSTICS

The study employs number of diagnostic test to test whether the selected model is adequate and efficient by testing the residuals of the model’s OLS estimated Vector Autoregressive (VAR) Model form. VAR model consist of a set of K endogenous variables $y_t = (y_{1t}, \dots, y_{kt}, \dots, y_{Kt})$ for $k = 1, \dots, K$. The VAR process is then defined as:

$$y_t = A_1 y_{t-1} + \dots + A_p y_{t-p} + C D_t + \mu_t \quad (3.48)$$

where A_i are $(K \times K)$ coefficient matrices for $i = 1, \dots, p$ and μ_t is a K dimensional white noise process with time-invariant positive definite covariance matrix $E(\mu_t \mu_t') = \Sigma \mu$. C denotes the coefficient matrix of potentially determined regressors with dimension $(K \times M)$,

and D_t is an $(M \times 1)$ column vector holding the appropriate deterministic regressors (Pfaff, 2008). The diagnostic tests employed in this study are serial correlation tests, White's heteroscedasticity test and normality test using the Jarque-Bera test. The VAR model is said to be very sensitive to assumptions of serial correlation in the residuals, while is not affected by heteroscedasticity and normality assumptions in the residuals. Therefore, the study will put emphasis in trying to correct serial correlation residuals by adding lags in the model until the AR disturbances in the residuals are uncorrelated.

In addition to present the ARCH disturbances, it is extensively known that financial data also has the problem of residuals being not normally distributed (negatively skewed, positively skewed) (Balarezo, 2010). Since there is not much that can be done to correct these problems of heteroscedasticity and normality in the residuals, and because they do not have a major effect in the VAR model, the study accepts them the way they are. The tests are discussed in the following subtitles.

3.13.1 Serial correlation test

The efficiency of the OLS is investigated by testing the model for serial correlation. Breusch-Godfrey (1980) Lagrange Multiplier (LM) test is used to test for serial correlation. The Breusch-Godfrey (BG) LM test can be applied regardless whether lagged dependent variables are included or not. The hypothesis for BG LM test is given as:

$$H_0: \rho_1 = \dots = \rho_h = 0$$

versus

$$H_1: \rho_i \neq 0 \quad \text{for } i = 1, 2, \dots, h$$

The BG LM test allows testing of the association between $\hat{\mu}_i$ and the numerous of its lagged values at the same time. The test is a more general test for serial correlation up to the r^{th} order. The BG LM test is based upon the following auxiliary regressions:

$$\hat{\mu}_t = A_1 y_{t-1} + \dots + A_p y_{t-p} + CD_t + \rho_1 \hat{\mu}_{t-1} + \dots + \rho_h \hat{\mu}_{t-h} + \varepsilon_t \quad (3.49)$$

The test statistic is given by:

$$LM_h = T[K - tr(\tilde{\Sigma}_R^{-1}\tilde{\Sigma}_e)] \quad (3.50)$$

where $\tilde{\Sigma}_R$ and $\tilde{\Sigma}_e$ are the residual covariance matrix of the restricted and unrestricted models respectively. The test statistic LM_h is distributed as chi square $\chi^2(hK^2)$. The null hypothesis is rejected if the probability in the serial correlation test statistic is less than the 5% level of significance (Pfaff, 2008).

3.13.2 Heteroscedasticity test

The heteroscedasticity test is effected by applying the Lagrange Multiplier (LM) in order to test for the existence or absence of heteroscedasticity in the residuals. The LM test was proposed by Engle (1982). The LM test is applied to the residuals of the cointegrating relationship to test the null hypothesis of homoscedasticity. The hypothesis is stated as follows:

H_0 : The residuals are homoscedastic

versus

H_1 : The residuals are heteroscedasticity

The LM test statistic is given by:

$$LM_E = nR^2 \quad (3.51)$$

where n is the number of observations, and R^2 is the coefficient of determination of the augmented residual regression. The coefficient of determination is given by:

$$R^2 = \frac{\text{regression sum of squares}}{\sum_{i=1}^n \hat{\alpha}_i^2} \quad (3.52)$$

$$R^2 = \frac{\sum_{i=1}^n (X_i \hat{\beta})^2}{\sum_{i=1}^n \hat{\alpha}_i^2} \quad (3.53)$$

$$R^2 = \frac{\hat{\beta}^T X^T X \hat{\beta}}{\hat{\alpha}^T \hat{\alpha}} \quad (3.54)$$

but

$$\hat{\beta} = (X^T X)^{-1} X^T \hat{\alpha}, \quad (3.55)$$

therefore

$$R^2 = \frac{\hat{\alpha}^T X (X^T X)^{-1} X^T \hat{\alpha}}{\hat{\alpha}^T \hat{\alpha}} \quad (3.56)$$

where $X_i = (x_{1i}, x_{2i})$, $X^T = (X_1^T, \dots, X_n^T)$ and $\hat{\alpha} = (\hat{\alpha}_1, \dots, \hat{\alpha}_n)^T$. The null hypothesis is rejected if the probability of the LM_E test is less than the 5 percent level of significance and it is concluded that there is heteroscedasticity. If there is a presence of heteroscedasticity, then the data is best modelled by Bollerslev's (1986) generalised autoregressive conditional heteroscedastic (GARCH) or autoregressive conditional heteroscedastic (ARCH) model.

3.13.3 Normality test

Lastly, the study tests whether or not residuals follow a normal distribution. The Jarque-Bera (JB) (1980) normality test, skewness and kurtosis tests are applied to test for normality. The JB test computes the difference in kurtosis and skewness of a variable compared to those of the normal distribution (Jarque and Bera, 1980). According to literature, financial returns series are hardly normally distributed. Tsay (2005:10) indicated that "financial time series are high-peaked, fat-tailed distributions of financial returns are common knowledge amongst practitioners." Cont (2001) also indicated that normal distribution is not possible in most of the financial time series data. The null and alternative hypothesis are stated as:

H_0 : The variable is normally distributed

versus

H_1 : The variable is not normally distributed

The JB test statistic is given as follows:

$$JB = \frac{T}{6} \left[s^2 s + \frac{1}{4} ((k - 3)^T (k - 3)) \right] \quad (3.57)$$

where T is the sample size, s and k are sample measures of skewness and kurtosis respectively. The JB test follows a $\chi^2(2k)$ and skewness and kurtosis tests are distributed $\chi^2(k)$, where K is

the number of endogenous variables. The formula for sample skewness and kurtosis are given by equation (3.58) and (3.59) respectively.

$$S = \frac{\sum_{t=1}^T (x_t - \bar{x})^3 / T - 1}{s^3} \quad (3.58)$$

$$k = \frac{\sum_{t=1}^T (x_t - \bar{x})^4 / T - 1}{s^4} \quad (3.59)$$

Under the assumption for normality, the coefficient of skewness and excess kurtosis are expected to be 0 and 3 respectively. The null hypothesis is rejected if the p-value is less than 5% level of significance.

3.14 CAUSALITY TEST

The cointegration technique initiated by Engle and Granger (1987), Granger (1986), and Hendry (1986) made a substantial contribution in the direction of testing Granger causality. The Granger technique tests whether or not adding lags of one variable have a predictive power for another variable. The ECM can be used to model the long run and short run Granger causal flows of the cointegrated share price series (Alexander, 2008). The causal relationship between two stock prices may be unidirectional, bidirectional or independent. According to Alexander (1993:2043) and Wakeford (2004:11), “the ECM specification is then used for the causality tests as it ensures stationarity of the variables.” The hypotheses tested is stated as follows:

$$H_0: \beta_{11}^1 = \beta_{12}^2 = \dots = \beta_{1i}^i = \gamma_1 = 0$$

$$H_1: \beta_{11}^1 \neq \beta_{12}^2 \neq \dots \neq \beta_{1i}^i \neq \gamma_1 = 0$$

$$H_0: \beta_{21}^1 = \beta_{22}^2 = \dots = \beta_{2i}^i = \gamma_2 = 0$$

$$H_1: \beta_{21}^1 \neq \beta_{22}^2 \neq \dots \neq \beta_{2i}^i \neq \gamma_2 = 0$$

The first null hypothesis indicates that B does not Granger cause A , the second null hypothesis indicates that A does not Granger cause B . The Granger causality test is represented by the follows equations:

$$\Delta A_t = \alpha_1 + \sum_{i=1}^m \beta_{11}^i \Delta A_{t-i} + \sum_{i=1}^m \beta_{12}^i \Delta B_{t-i} + \gamma_1 z_{t-1} + \varepsilon_{1t} \quad (3.60)$$

$$\Delta B_t = \alpha_2 + \sum_{i=1}^m \beta_{21}^i \Delta A_{t-i} + \sum_{i=1}^m \beta_{22}^i \Delta B_{t-i} + \gamma_2 z_{t-1} + \varepsilon_{2t} \quad (3.61)$$

where ΔA_t and ΔB_t represent the first difference of the time series under study, α_1 and α_2 are constants, and ε_{1t} and ε_{2t} are white noise terms. In equation (3.60), B Granger causes A if the joint significance test of all variables with lagged B and γ_1 are different from zero. This means that the past values of B helps to predict the current or future values of A better than the past values of A alone (Alexander, 2008). While in equation (3.61), A Granger causes B if the joint significance test of all variables with lagged A and γ_2 are different from zero. If both γ_1 and γ_2 are equal to zero, there is no presence of long run Granger causality and the model is neither an EC nor a cointegration (Enders, 1995). The hypothesis are tested using an F test presented in the following formula:

$$F_{calc} = \frac{(SSR_R - SSR_{UR})/p}{(SSR_{UR}/n-k-1)} \quad (3.62)$$

where SSR_R and SSR_{UR} are sum of squares for residuals of the restricted and unrestricted models respectively, p denotes number of lagged terms, k is the number of parameters estimated in the unrestricted model and n is the number of observations. The null hypothesis is rejected if the p-value of the F_{calc} is less than 5% level of significance (Brandt and Williams, 2006).

3.15 STATISTICAL ARBITRAGE TRADING STRATEGY

According to Meki (2012:57), “the pairs trading strategy is a statistical arbitrage technique implemented mainly to exploit short run price deviations from a pair of stocks believed to share long run price equilibrium.” Battal (2009:18) stated that “cointegration based pair trading is implemented by determining the long run relationships between the historical price series of two financial assets, and taking the advantage of the short run deviations from this relationship.” This arbitrage strategy was designed to exploit short term deviations from a long run equilibrium between pairs of stocks. The pair trading is a statistical arbitrage trading strategy that enables traders or investors to make a profit from almost any market conditions. The pairs trading strategy used in this study is based on the assumption that a linear combination of stock prices reverts to a long run equilibrium (cointegration) and a trading rule can be constructed to take advantage of the temporary deviations. Assuming that long run equilibrium relation is given by a linear function $\beta'X_t = 0$ of a paired log prices in vector X_t , process of deviation (mispricing) defined as $Y_t = \beta'X_t$ should be stationary, autoregressive process (Chan, 2011).

Pairs trading matches a long position with a short position in a pair of highly correlated and cointegrated pairs of stock. Pairs traders wait for weakness in the correlation or divergence from the long term equilibrium in cointegration, and then go long on the underperforming stock while simultaneously going short on the over performing stock (Jacobs and Levy, 1993). Pairs-traders would close the positions as the relationship returns to its statistical normal. The profit of the strategy is derived from the difference in price change between the two stocks, rather than from the direction in which each moves. The difference in prices is called spread. The spread represent the degree of mispricing of one asset relative to the other. To profit from relative mispricing, the trade is opened by buying the underperforming stock and sell (short) the relative stock that is over-performing. The spread is calculated using the following equation:

$$\varepsilon_t = \ln(P_{y_t}) - \beta_t \ln(P_{x_t}) \quad (3.63)$$

where ε_t represents the value of the spread at time t , P_{x_t} is the price of stock X at time t , P_{y_t} is the price of stock Y at time t and β_t is the adaptive coefficient beta at time t . After calculating the spread using equation (3.63), then the spread is normalized by subtracting its mean and dividing by its standard deviation. When the spread reaches the predetermined threshold, the degree of mutual mispricing between the paired stocks has diverged far from its historical mean. This means that trade can be opened.

Steps to follow when implementing pairs trading are as follows. First step is to formulate pairs using the correlation and cointegration methods discussed in previous sections. Once the pairs are formulated, the next step is to calculate price ratio. Price ratio is often referred to as relative performance line. The price ratio is computed by dividing the price of one stock by its corresponding pair⁴. This ratio represent the average spread between the paired stocks. The ratio is also referred to as the volatility boundaries. Third step will be to spot which spikes to trade on and when to exit.

When P_t^A and P_t^B are cointegrated, ε_t is stationary. This implies that

⁴ Choosing which stock price to be the denominator or numerator is arbitrary to the author.

$$\lim_{T \rightarrow \infty} \left(\sqrt{\frac{1}{T} \sum_{t=0}^T \varepsilon_t} \right) = \mu \quad (3.64)$$

and

$$\lim_{T \rightarrow \infty} \left(\sqrt{\frac{1}{T-1} \sum_{t=0}^T (\varepsilon_t - \mu)^2} \right) = SD \quad (3.65)$$

where μ is the long run equilibrium for the relative price of P_t^A and P_t^B , it is believed that at some point in time ε_t will revert to μ . An entry is estimated by using the standard deviation (SD) of the paired stock price ratio. Pelletier (2007:2) defined z-score as “the distance measured in terms of standards deviations, between the spread and its mean.” The dimensionless z-score to measure the distance between the long term mean in units of long term standard deviation is given by:

$$Z_t = \frac{\varepsilon_t - \mu_\varepsilon}{\sigma_\varepsilon} \quad (3.66)$$

where μ_ε is the 250-day mean value, ε_t is the value of the price spread at time t and σ_ε denotes the standard deviation of the trading signal. Based on the historical prices, the entry points are set at ± 2 SD from the one-year moving average. This means that when the price ratio is above (below) 2 SD, there are profitable opportunities to trade when going short (long) on the numerator and long (short) on the denominator. An investor may buy (long) the spread at a certain time period t when the relative mispricing reaches -2 SD and sell (short) when spread reaches 2 SD. More precisely, an investor would buy (long) spread at time t when the relative mispricing reaches -2 SD from the mean and sell (short) the spread at time $t+i$:

$$Y_t - \hat{\beta}X_t = \hat{\mu} - 2SD \quad (3.67)$$

$$Y_{t+i} - \hat{\beta}X_{t+i} = \hat{\mu} \quad (3.68)$$

On the other hand, an investor would sell (short) the spread at time t when it reaches 2 SD above the mean and buy (long) the spread at time $t + i$:

$$Y_t - \hat{\beta}X_t = \hat{\mu} + 2SD \quad (3.69)$$

$$Y_{t+i} - \hat{\beta}X_{t+i} = \hat{\mu} \quad (3.70)$$

where $\hat{\mu}$ is referred to as the closing threshold.

According to Gatev et al. (2006:804), “traders generally use a rule of thumb, namely the two standard deviation rule, in entering a pairs trade (-A, B) or (A, -B)”. Therefore, this study is going to use the rule of thumb as proposed by Gatev et al. (2006). Last step pairs trading will be to implement pairs trading strategy based on the above mentioned steps. The trading signal can be summarised as:

Buy to open if $\varepsilon_t < -2.00$

Sell to open if $\varepsilon_t > 2.00$

Close position if ε_t crosses μ_t

Pairs trading strategy is implemented using sample period of 250 trading days.

3.16 CHAPTER SUMMARY

This chapter presented the methodology applied to four sectors in the implementation of the statistical arbitrage technique on the JSE Top 40 Index Option. Preliminary data analysis gives a clear understanding of the characteristics of the financial time series data used in the study. According to literature, financial time series is characterised by stylised facts and the most common stylised facts are no presence of autocorrelation, flat tails and gain/loss asymmetry. The stylised facts of the data used in this study are derived from the descriptive statistics of the log transformed data. The graphical and formal approach of testing for stationarity were discussed. Simple linear regression and Engle and Granger cointegration techniques are going to be used to test the relationship between each paired stock prices. Furthermore, the model diagnostics and Granger causality techniques are also going to be applied in this study. Finally, statistical arbitrage trading technique will be implemented on the cointegrated pairs. Chapter 4 presents data analysis and interpretation of the results.

CHAPTER 4

DATA ANALYSIS AND INTERPRETATION OF THE RESULTS

4.1 INTRODUCTION

This chapter focuses on data analysis and the interpretation of the results. The data is presented first in graphical format in order to get some idea about the underlying data generating process governing each series. It is important to graphically present the data before the actual implementation of the statistical arbitrage method on JSE Top 40 Index Option as this reveals the hidden patterns in the data and provides direction about the models that are derived to capture the essence of the data fit.

The analysis is divided into four sections. Section 4.2.1 presents the analysis on the financial sector, Section 4.3.1 presents analysis on the industrial sector, Section 4.4.1 presents analysis on the mining sector and the last Section, 4.5.1, presents the analysis on the retail sector. For every Section, the following concepts are covered: preliminary data analysis, correlation analysis, test for stationarity using the Augmented Dickey–Fuller (ADF) test and Kwiatkowski, Phillips, Schmidt and Shin (KPSS) test, the establishment of the relationship between index pairs, the determination of the boundaries for regression coefficients, pairs trading using cointegration technique and the implementation of the statistical arbitrage trading strategy.

4.2 RELIMINARY DATA ANALYSIS

4.2.1 Analysis on financial sector

Preliminary analysis gives characteristics of the financial sector. The descriptive measures used are the mean, median, maximum (max), minimum (min), standard deviation, skewness, kurtosis, Jarque-Bera (JB) test for normality together with their p-values and Ljung-Box (LB) statistics computed for 15 lags together with the p-value. The results are presented in Table 4.1.1.

Table 4.1.1: Descriptive statistics for log transformed financial sector

<i>Index</i>	<i>LBGA</i>	<i>LBVT</i>	<i>LDSY</i>	<i>LFSR</i>	<i>LINL</i>	<i>LIPL</i>	<i>LNED</i>	<i>LOML</i>	<i>LREI</i>	<i>LRMH</i>	<i>LSBK</i>	<i>LSLM</i>
Mean	9.608	9.956	8.767	7.992	8.809	9.644	9.787	7.811	7.453	8.321	9.380	8.314
Median	9.570	9.979	8.724	7.974	8.717	9.769	9.804	7.823	7.410	8.280	9.339	8.372
Max	9.901	10.487	9.646	8.666	9.413	10.026	10.207	8.427	8.073	8.894	9.777	9.000
Min	9.393	9.408	8.016	7.475	8.324	8.955	9.369	7.193	6.963	7.742	9.108	7.696
Std. Dev.	0.117	0.317	0.473	0.365	0.308	0.281	0.230	0.372	0.330	0.309	0.154	0.388
Skewness	0.546	-0.073	0.145	0.290	0.466	-0.570	-0.056	-0.028	0.189	0.184	0.647	0.052
Kurtosis	2.190	1.543	1.625	1.738	1.823	1.865	1.660	1.510	1.595	1.966	2.462	1.572
JB test	115.514	133.956	123.337	120.655	140.880	161.858	113.047	138.922	132.427	75.345	122.866	128.0
Prob.	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
LB (15)	20377	21870	21959	22072	21895	21406	21840	21953	21928	21850	21400	22047
Prob.	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
N	1500	1500	1500	1500	1500	1500	1500	1500	1500	1500	1500	1500

The results in Table 4.1.1 revealed that the log of Barclays Africa Group Ltd (LBGA), Discovery Ltd (LDSY), FirstRand Limited (LFSR), Investec Limited (LINL), Reinet Investments (LREI), RMB Holdings (LRMH), Standard Bank Group (LSBK) and Sanlam (LSLM) are positively skewed. This means that most of the stock prices are concentrated to the left. Bidvest Group (LBVT), Imperial Holdings (LIPL), Nedbank Group (LNED), (LOML) and Old Mutual (LOML) are negatively skewed. This means that most of the stock prices are concentrated to the right. Performance based on the skewness illustrates that for negative skewness five sectors were performing better during the selected period and the positive skewness reports contrary findings.

The kurtosis values of all the variables are less than 3, which means that variables are platykurtically distributed. The p-values of the JB test are all less than 5% level of significance. Therefore, it is concluded that all the variables are not normally distributed, that is, they are not symmetrical as expected. The data further revealed that there is a presence of autocorrelation up to the 15th lag. The three statistics reported confirm the stylised facts about financial data and this necessitates the implementation of the statistical arbitrage on the data.

4.2.2 Graphical presentation of financial sector

The graphical presentation of the data gives some idea about the underlying data generating process governing each stock price. Understanding these processes helped in determining what formal tests need to be administered in the subsequent analysis paving way to address the main goal of this study. The variables under financial sector are presented in Figure 4.1.1.



Figure 4.1.1: Time series plot of the historical volatility of financial sector at level

Figure 4.1.1 depicts the time series plot of the 12 companies classified under financial sector category. The volatility patterns of Bidvest Group (BVT) and Discovery Ltd (DSY) seem to be increasing progressively over time compared to other indexes. This means that over the time specified in the study, the stock prices for both BVT and DSY were increasing. Sanlam (SLM) and nine other indexes around it show a slight trend. In order to understand the influential forces behind each series, the data was smoothed by logarithm transformation. The stock prices revealed no sign of reversion around the mean value, and this indicates that the data is nonstationary at level. Therefore, log transformation was applied to financial sector variables to stabilise the variables and the results are presented in Figure 4.1.2.

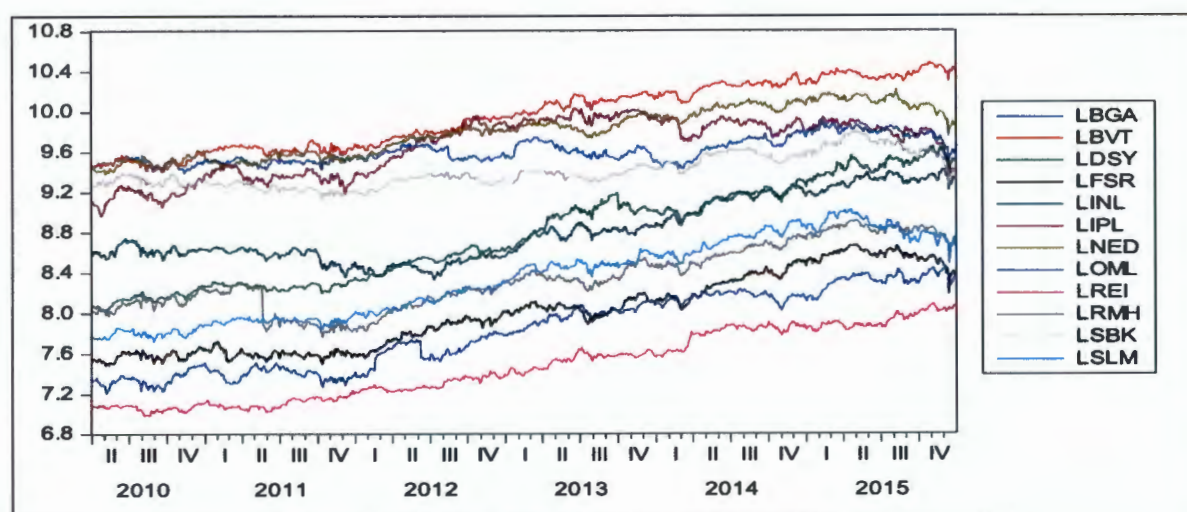


Figure 4.1.2: Time series plot of the financial sector at logarithm

Figure 4.1.2 illustrates that the log transformed financial sector does not show any sign of mean reversion. Therefore, it is concluded that the logarithm of the financial sector is nonstationary. This means that the data must be differenced. The following Figure 4.1.3 depicts the log differenced series.

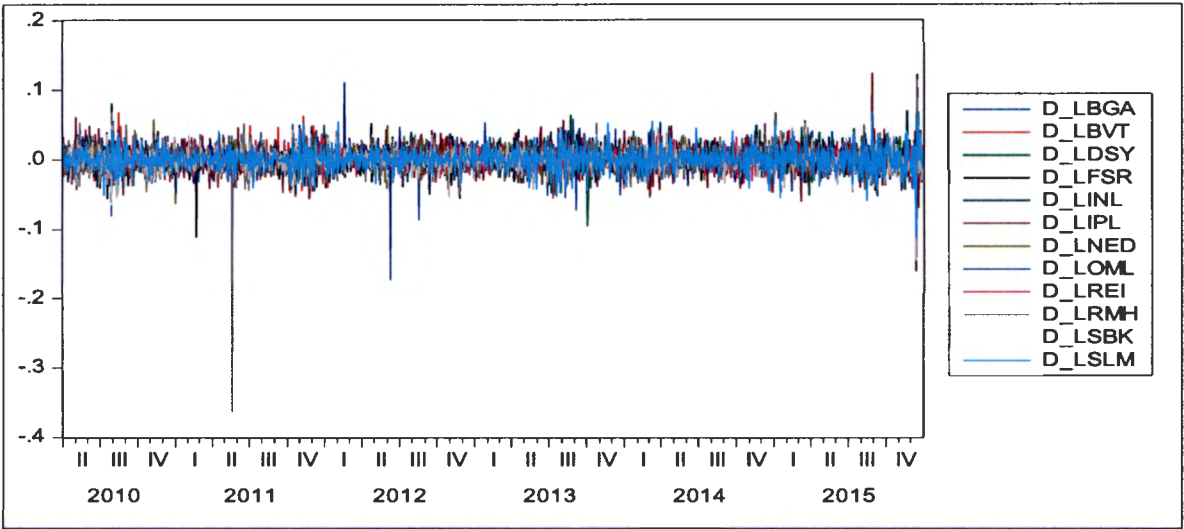


Figure 4.1 3: Time series plot of the financial sector at logarithm difference

Figure 4.1.3 depicts that the first log differenced indexes fluctuate around the mean value and the variance is relatively stable. Since the data was differenced once, it is concluded that the data is integrated of order one, $I(1)$.

Presented next are the pairwise correlations for the 12 financial sector companies. As discussed earlier, a bench mark of 0.90 was referred to when interpreting the results. Only the pairs with the highest (minimum 0.90) correlation coefficient were selected and considered for further analyses.

Table 4.1.2: Pairwise correlation of the financial sector

<i>Correlation</i>	<i>BGA</i>	<i>BVT</i>	<i>DSY</i>	<i>FSR</i>	<i>INL</i>	<i>IPL</i>	<i>NED</i>	<i>OML</i>	<i>REI</i>	<i>RMH</i>	<i>SBK</i>	<i>SLM</i>
BGA	1											
BVT	0.773	1										
DSY	0.798	0.972	1									
FSR	0.872	0.954	0.964	1								
INL	0.768	0.907	0.938	0.941	1							
IPL	0.554	0.755	0.650	0.649	0.467	1						
NED	0.830	0.949	0.909	0.956	0.858	0.789	1					
OML	0.758	0.983	0.972	0.950	0.919	0.718	0.941	1				
REI	0.761	0.977	0.971	0.949	0.940	0.642	0.921	0.971	1			
RMH	0.828	0.925	0.942	0.975	0.952	0.590	0.912	0.924	0.921	1		
SBK	0.852	0.852	0.861	0.934	0.896	0.567	0.904	0.869	0.845	0.935	1	
SLM	0.830	0.968	0.942	0.973	0.910	0.731	0.972	0.955	0.946	0.942	0.907	1

The pairwise correlation presented in Table 4.1.2 indicates that thirty nine pairs of stocks were selected. The index pairs with a correlation coefficient equal to or higher than 0.90 (bold) are further analysed. The choice of using 90% as a cut-off point was discussed in Chapter 3. The thirty nine pairs of stock were then tested for stationarity to avoid spurious regression. Formal tests for stationarity are presented in Table 4.1.3.

4.2.3 Test for stationarity

As mentioned in the preceding chapter, stationarity is examined by testing the ten stock indexes for the presence of unit root. The two tests employed are the ADF test and the KPSS test. The stationarity test is only applied to the selected significant indexes. The results are presented in Table 4.1.3.

Table 4.1.3: Unit root tests of the financial sector

<i>Index</i>	<i>ADF test statistic</i>	<i>p-value</i>	<i>KPSS test statistic</i>	<i>p-value</i>
BVT	-0.463	0.896	4.731	0.000
LBVT	-0.775	0.825	4.748	0.000
ΔLBVT	-10.612	0.000**	0.036	0.098**
DSY	-0.102	0.947	4.479	0.000
LDSY	-0.556	0.878	4.706	0.000
ΔLDSY	-10.962	0.000**	0.037	0.014
FSR	-0.815	-0.815	4.358	0.000
LFSR	-0.894	0.791	4.576	0.000
ΔLFSR	-18.049	0.000**	0.098	0.200**

<i>Index</i>	<i>ADF test statistic</i>	<i>p-value</i>	<i>KPSS test statistic</i>	<i>p-value</i>
INL	-0.021	0.956	3.963	0.000
LINL	-0.340	0.917	3.951	0.000
Δ LINL	-41.338	0.000**	0.129	0.210**
NED	-1.497	0.535	4.475	0.000
LNED	-1.611	0.477	4.522	0.000
Δ LNED	-18.187	0.000**	0.204	0.476**
OML	-0.428	0.902	4.629	0.000
LOML	-0.680	0.850	4.663	0.000
Δ LOML	-40.369	0.000**	0.039	0.089**
REI	1.127	0.998	4.617	0.000
LREI	0.296	0.978	4.741	0.000
Δ LREI	-30.195	0.000**	0.125	0.029
RMH	-1.010	0.752	4.027	0.000
LRMH	-1.081	0.725	4.001	0.000
Δ LRMH	-30.015	0.000**	0.081	0.407**
SBK	-1.615	0.475	3.532	0.000
LSBK	-1.679	0.442	3.621	0.000
Δ LSBK	-31.182	0.000**	0.200	0.848**
SLM	-0.998	0.756	4.526	0.000
LSML	-1.127	0.707	4.682	0.000
Δ LSML	-29.006	0.000**	0.101	0.111**

*Note: The ADF and KPSS critical values for the 1%, 5% and 10% significance levels are -3.4345, -2.8633, -2.5677 and 0.7314, 0.4630, 0.3470 respectively. *, ** and *** denotes rejection of the null hypothesis at 1% and 5% and 10% respectively. Lag length was selected by AIC for ADF. Δ and L denotes first difference and log transformation respectively.*

Table 4.1.3 shows that all the original series were nonstationary at level since all the p-values for ADF test are more than 5% level of significance and all the p-values of KPSS are less than 5% level of significance. The same conclusions were also reached for all the log transformed variables for both ADF and KPSS test. The log differenced data revealed that the null hypothesis for ADF unit root was rejected for all the variables, meaning that all the variables became stationary after first log difference. The p-values for KPSS test also revealed that most of the variables were stationary at log difference except for Δ LDSY and Δ LREI. This conclusion was reached using ADF test since all the variables were stationary at first log difference. Therefore, it is concluded that all the variables are integrated to order 1, $I(1)$. The next step is to establish the relationship between the index pairs as discussed in Chapter 3.

4.2.4 Relationship between the index pairs

This section establishes the relationship between the paired stock indexes. Each and every pair of the selected indexes are regressed using the OLS technique. The OLS regression is used to establish the relationship between each of the thirty nine pairs of highly correlated indexes. The regression equations for $\Delta LDSY$ - $\Delta LBVT$, $\Delta LFSR$ - $\Delta LBVT$, $\Delta LINL$ - $\Delta LBVT$, $\Delta LNED$ - $\Delta LBVT$, $\Delta LOML$ - $\Delta LBVT$, $\Delta LREI$ - $\Delta LBVT$, $\Delta LRMH$ - $\Delta LBVT$ and $\Delta LSLM$ - $\Delta LBVT$ pairs are as follows:

$$R_{DSY\ i} = \hat{\beta}_1 + \hat{\beta}_2 R_{BVT\ i} + \hat{\mu}_i \quad (4.1.1)$$

$$R_{FSR\ i} = \hat{\beta}_1 + \hat{\beta}_2 R_{BVT\ i} + \hat{\mu}_i \quad (4.1.2)$$

$$R_{INL\ i} = \hat{\beta}_1 + \hat{\beta}_2 R_{BVT\ i} + \hat{\mu}_i \quad (4.1.3)$$

$$R_{NED\ i} = \hat{\beta}_1 + \hat{\beta}_2 R_{BVT\ i} + \hat{\mu}_i \quad (4.1.4)$$

$$R_{OML\ i} = \hat{\beta}_1 + \hat{\beta}_2 R_{BVT\ i} + \hat{\mu}_i \quad (4.1.5)$$

$$R_{REI\ i} = \hat{\beta}_1 + \hat{\beta}_2 R_{BVT\ i} + \hat{\mu}_i \quad (4.1.6)$$

$$R_{RMH\ i} = \hat{\beta}_1 + \hat{\beta}_2 R_{BVT\ i} + \hat{\mu}_i \quad (4.1.7)$$

$$R_{SLM\ i} = \hat{\beta}_1 + \hat{\beta}_2 R_{BVT\ i} + \hat{\mu}_i \quad (4.1.8)$$

The relationship between the first eight pairs is represented in Figure 4.1.4 respectively.

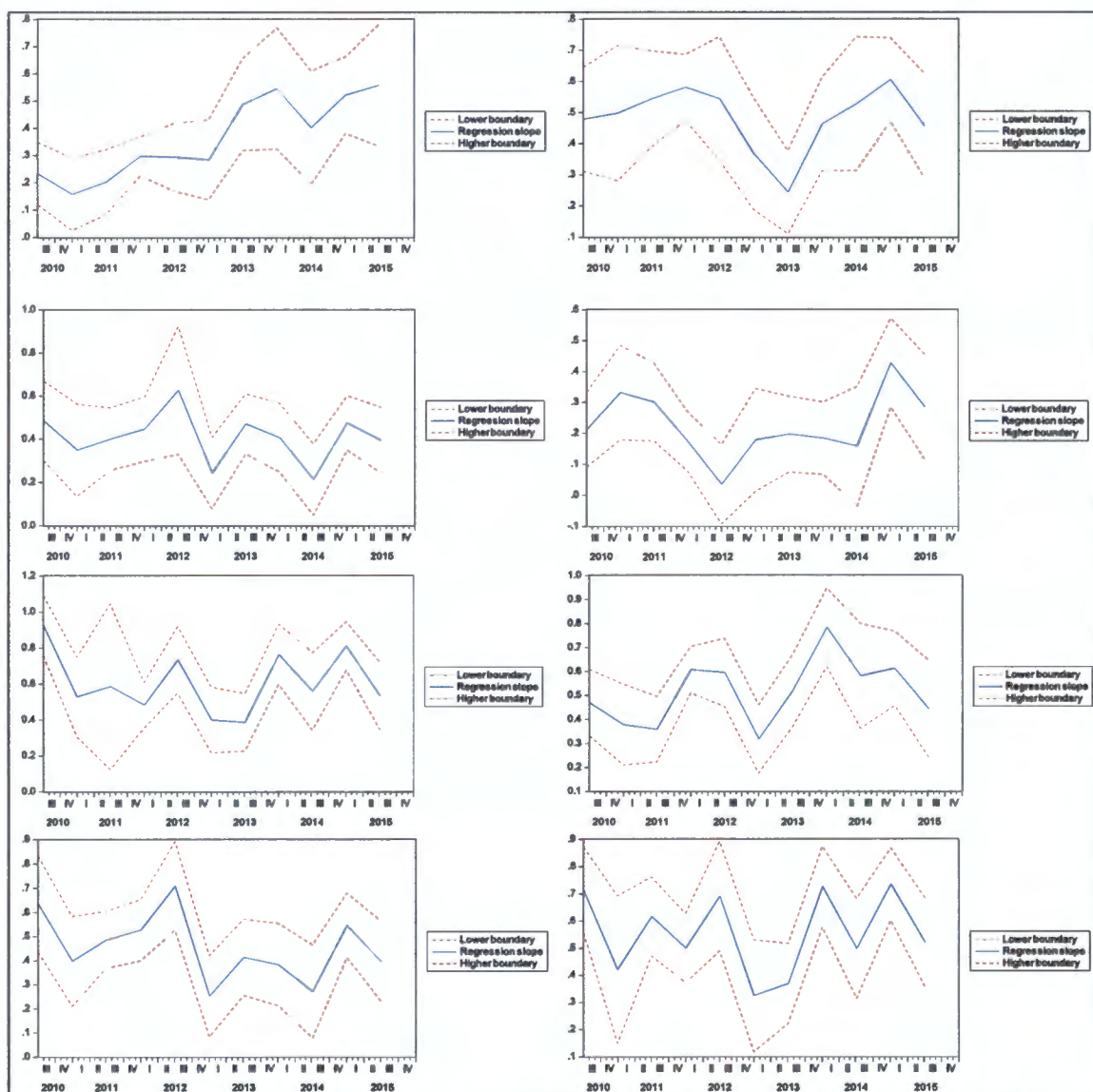


Figure 4.1.4: Beta boundaries for $\Delta\text{LDSY}-\Delta\text{LBVT}$, $\Delta\text{LFSR}-\Delta\text{LBVT}$, $\Delta\text{LINL}-\Delta\text{LBVT}$, $\Delta\text{LNED}-\Delta\text{LBVT}$, $\Delta\text{LOML}-\Delta\text{LBVT}$, $\Delta\text{LREI}-\Delta\text{LBVT}$, $\Delta\text{LRMH}-\Delta\text{LBVT}$ and $\Delta\text{LSLM}-\Delta\text{LBVT}$

Figure 4.1.4 presents the estimates of $\hat{\beta}_2$ values from rolling 125 day regressions starting from 07/05/2010 to 07/06/2015. The graph also depicts the calculated 95% confidence interval. The rolling regression slope falls between the lower and higher boundaries. This means that there is a 95% likelihood that a future index options would fall within the estimated interval. The rolling coefficients were tested for significance and the results are presented in Table 4.1.4.

Table 4.1.4: Beta boundaries for $\Delta LDSY-\Delta LBVT$, $\Delta LFSR-\Delta LBVT$, $\Delta LINL-\Delta LBVT$, $\Delta LNED-\Delta LBVT$, $\Delta LOML-\Delta LBVT$, $\Delta LREI-\Delta LBVT$, $\Delta LRMH-\Delta LBVT$ and $\Delta LSLM-\Delta LBVT$

Date	$\Delta LDSY-\Delta LBVT$			$\Delta LFSR-\Delta LBVT$			$\Delta LINL-\Delta LBVT$			$\Delta LNED-\Delta LBVT$		
	$\hat{\beta}_2$	t-test	p-value	$\hat{\beta}_2$	t-test	p-value	$\hat{\beta}_2$	t-test	p-value	$\hat{\beta}_2$	t-test	p-value
07/05/2010	0.233	4.064	0.000	0.715	9.070	0.000	0.636	6.392	0.000	0.478	5.704	0.000
12/31/2010	0.157	2.370	0.019	0.422	3.106	0.002	0.398	4.278	0.000	0.498	4.578	0.000
07/05/2011	0.202	3.390	0.001	0.618	8.492	0.000	0.487	8.238	0.000	0.543	7.035	0.000
01/03/2012	0.297	8.325	0.000	0.501	7.802	0.000	0.528	8.300	0.000	0.581	11.007	0.000
07/03/2012	0.293	4.671	0.000	0.692	6.879	0.000	0.710	7.795	0.000	0.544	5.469	0.000
01/02/2013	0.284	3.869	0.000	0.325	3.152	0.002	0.253	2.994	0.003	0.367	4.092	0.000
07/03/2013	0.488	5.740	0.000	0.370	5.024	0.000	0.414	5.220	0.000	0.244	3.669	0.000
01/02/2014	0.545	4.900	0.000	0.727	9.791	0.000	0.384	4.491	0.000	0.463	6.164	0.000
07/07/2014	0.402	3.870	0.000	0.499	5.417	0.000	0.273	2.822	0.006	0.528	4.915	0.000
01/05/2015	0.522	7.393	0.000	0.736	11.102	0.000	0.547	8.212	0.000	0.606	9.036	0.000
07/06/2015	0.559	4.940	0.000	0.522	6.372	0.000	0.396	4.759	0.000	0.457	5.410	0.000
Date	$\Delta LOML-\Delta LBVT$			$\Delta LREI-\Delta LBVT$			$\Delta LRMH-\Delta LBVT$			$\Delta LSLM-\Delta LBVT$		
	$\hat{\beta}_2$	t-test	p-value	$\hat{\beta}_2$	t-test	p-value	$\hat{\beta}_2$	t-test	p-value	$\hat{\beta}_2$	t-test	p-value
07/05/2010	0.486	5.296	0.000	0.486	5.296	0.000	0.213	3.518	0.001	0.925	11.024	0.000
12/31/2010	0.350	3.287	0.001	0.350	3.287	0.001	0.332	4.347	0.000	0.529	4.783	0.000
07/05/2011	0.402	5.562	0.000	0.402	5.562	0.000	0.301	4.784	0.000	0.586	2.546	0.012
01/03/2012	0.448	6.001	0.000	0.448	6.001	0.000	0.173	3.614	0.000	0.483	7.638	0.000
07/03/2012	0.627	4.221	0.000	0.627	4.221	0.000	0.036	0.560	0.576	0.734	7.908	0.000
01/02/2013	0.244	2.974	0.004	0.244	2.974	0.004	0.179	2.157	0.033	0.400	4.386	0.000
07/03/2013	0.472	6.769	0.000	0.472	6.769	0.000	0.198	3.239	0.002	0.387	4.865	0.000
01/02/2014	0.408	5.114	0.000	0.408	5.114	0.000	0.184	3.142	0.002	0.765	9.119	0.000
07/07/2014	0.215	2.614	0.010	0.215	2.614	0.010	0.158	1.642	0.103	0.560	5.186	0.000
01/05/2015	0.476	7.574	0.000	0.476	7.574	0.000	0.428	5.945	0.000	0.812	12.046	0.000
07/06/2015	0.395	5.177	0.000	0.395	5.177	0.000	0.286	3.369	0.001	0.534	5.690	0.000

Table 4.1.4 above depicts the rolling coefficients together with their t-tests and p-values. Most of the estimated slope coefficients were found to be significantly different from zero at a 95% confidence level confirming the results presented in Figure 4.1.1. The p-values of most of the coefficients are less than 5% level of significance. Therefore the null hypothesis of $\hat{\beta}_2 = 0$ is rejected. This means that the paired variables are related. The relationship between the second eight pairs is presented below.

The regression equations for $\Delta\text{LFSR}-\Delta\text{LDSY}$, $\Delta\text{LINL}-\Delta\text{LDSY}$, $\Delta\text{LNED}-\Delta\text{LDSY}$, $\Delta\text{LOML}-\Delta\text{LDSY}$, $\Delta\text{LREI}-\Delta\text{LDSY}$, $\Delta\text{LRMH}-\Delta\text{LDSY}$, $\Delta\text{LSLM}-\Delta\text{LDSY}$ and $\Delta\text{LINL}-\Delta\text{LFSR}$ are as follows:

$$R_{\text{FSR } i} = \hat{\beta}_1 + \hat{\beta}_2 R_{\text{DSY } i} + \hat{\mu}_i \quad (4.2.1)$$

$$R_{\text{INL } i} = \hat{\beta}_1 + \hat{\beta}_2 R_{\text{DSY } i} + \hat{\mu}_i \quad (4.2.2)$$

$$R_{\text{NED } i} = \hat{\beta}_1 + \hat{\beta}_2 R_{\text{DSY } i} + \hat{\mu}_i \quad (4.2.3)$$

$$R_{\text{OML } i} = \hat{\beta}_1 + \hat{\beta}_2 R_{\text{DSY } i} + \hat{\mu}_i \quad (4.2.4)$$

$$R_{\text{REI } i} = \hat{\beta}_1 + \hat{\beta}_2 R_{\text{DSY } i} + \hat{\mu}_i \quad (4.2.5)$$

$$R_{\text{RMH } i} = \hat{\beta}_1 + \hat{\beta}_2 R_{\text{DSY } i} + \hat{\mu}_i \quad (4.2.6)$$

$$R_{\text{SLM } i} = \hat{\beta}_1 + \hat{\beta}_2 R_{\text{DSY } i} + \hat{\mu}_i \quad (4.2.7)$$

$$R_{\text{INL } i} = \hat{\beta}_1 + \hat{\beta}_2 R_{\text{FSR } i} + \hat{\mu}_i \quad (4.2.8)$$

The relationship between is represented in Figure 4.1.5 respectively

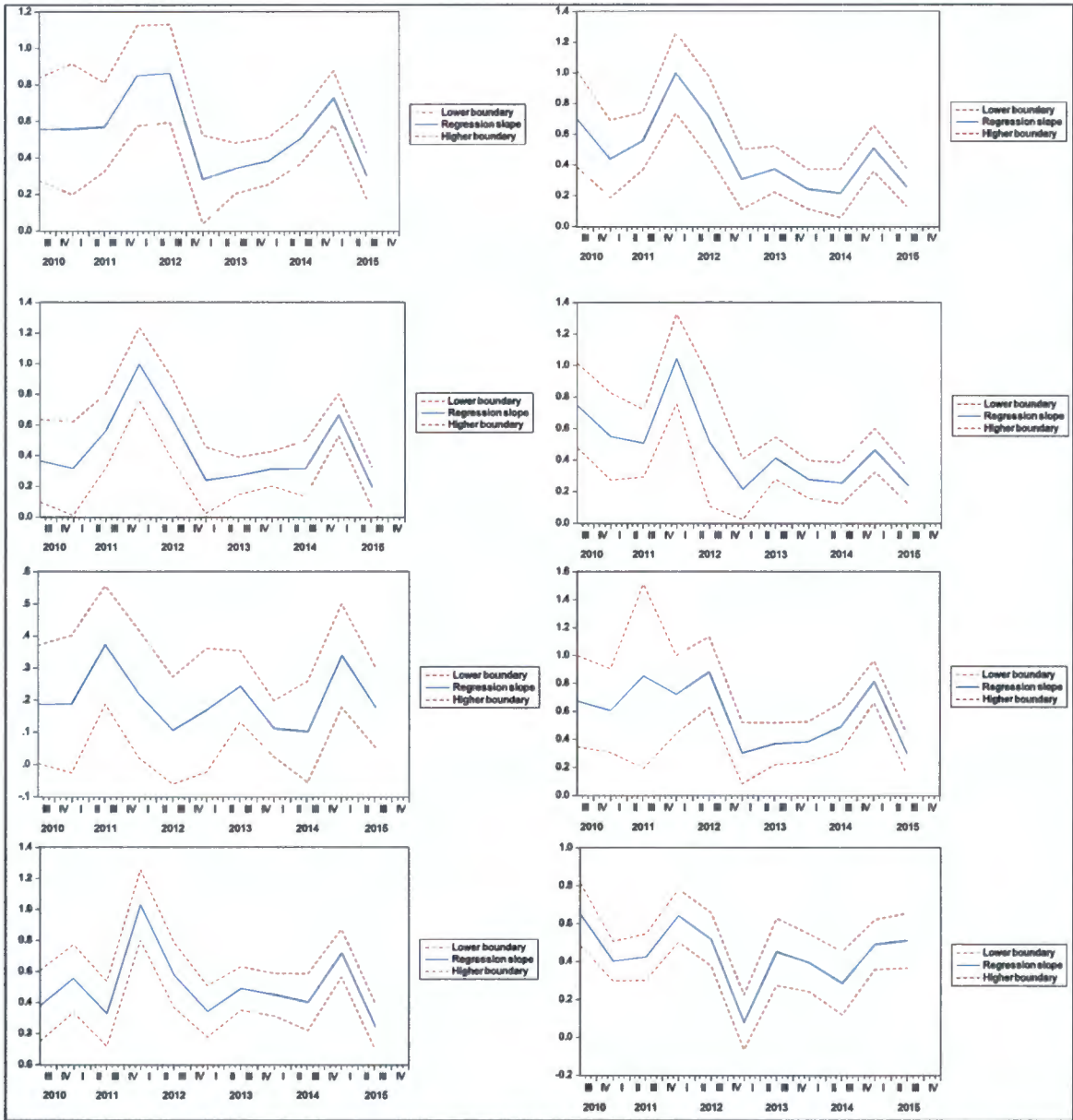


Figure 4.1.5: Beta boundaries for $\Delta\text{LFSR-}\Delta\text{LDSY}$, $\Delta\text{LINL-}\Delta\text{LDSY}$, $\Delta\text{LNED-}\Delta\text{LDSY}$, $\Delta\text{LOML-}\Delta\text{LDSY}$, $\Delta\text{LREI-}\Delta\text{LDSY}$, $\Delta\text{LRMH-}\Delta\text{LDSY}$, $\Delta\text{LSLM-}\Delta\text{LDSY}$ and $\Delta\text{LINL-}\Delta\text{LFSR}$

Figure 4.1.5 depicts the rolling regression slopes together with their 95% confidence limits. The rolling regression slopes fall within confidence bands. The regression slopes were tested for significance and the results are presented Table 4.1.5.

Table 4.1.5: Beta boundaries for $\Delta\text{LFSR}-\Delta\text{LDSY}$, $\Delta\text{LINL}-\Delta\text{LDSY}$, $\Delta\text{LNED}-\Delta\text{LDSY}$, $\Delta\text{LOML}-\Delta\text{LDSY}$, $\Delta\text{LREI}-\Delta\text{LDSY}$, $\Delta\text{LRMH}-\Delta\text{LDSY}$, $\Delta\text{LSLM}-\Delta\text{LDSY}$ and $\Delta\text{LINL}-\Delta\text{LFSR}$

<i>Date</i>	<i>$\Delta\text{LFSR}-\Delta\text{LDSY}$</i>			<i>$\Delta\text{LINL}-\Delta\text{LDSY}$</i>			<i>$\Delta\text{LNED}-\Delta\text{LDSY}$</i>			<i>$\Delta\text{LOML}-\Delta\text{LDSY}$</i>		
	$\hat{\beta}_2$	t-test	p-value	$\hat{\beta}_2$	t-test	p-value	$\hat{\beta}_2$	t-test	p-value	$\hat{\beta}_2$	t-test	p-value
07/05/2010	0.556	3.934	0.000	0.699	4.460	0.000	0.366	2.718	0.008	0.745	5.569	0.000
12/31/2010	0.556	3.096	0.002	0.439	3.489	0.001	0.317	2.079	0.040	0.550	3.984	0.000
07/05/2011	0.567	4.690	0.000	0.556	5.984	0.000	0.564	4.669	0.000	0.507	4.760	0.000
01/03/2012	0.849	6.165	0.000	0.998	7.575	0.000	0.998	8.376	0.000	1.043	7.326	0.000
07/03/2012	0.862	6.398	0.000	0.714	5.446	0.000	0.639	4.774	0.000	0.516	2.536	0.012
01/02/2013	0.282	2.334	0.021	0.305	3.151	0.002	0.241	2.232	0.027	0.216	2.255	0.026
07/03/2013	0.342	4.944	0.000	0.372	4.963	0.000	0.270	4.429	0.000	0.413	6.186	0.000
01/02/2014	0.382	5.950	0.000	0.243	3.780	0.000	0.315	5.559	0.000	0.277	4.648	0.000
07/07/2014	0.506	7.245	0.000	0.215	2.732	0.007	0.315	3.460	0.001	0.255	3.934	0.000
01/05/2015	0.727	9.789	0.000	0.507	6.789	0.000	0.666	9.667	0.000	0.463	6.743	0.000
07/06/2015	0.302	4.823	0.000	0.257	4.190	0.000	0.195	2.970	0.004	0.238	4.170	0.000
<i>Date</i>	<i>$\Delta\text{LREI}-\Delta\text{LDSY}$</i>			<i>$\Delta\text{LRMH}-\Delta\text{LDSY}$</i>			<i>$\Delta\text{LSLM}-\Delta\text{LDSY}$</i>			<i>$\Delta\text{LINL}-\Delta\text{LFSR}$</i>		
	$\hat{\beta}_2$	t-test	p-value	$\hat{\beta}_2$	t-test	p-value	$\hat{\beta}_2$	t-test	p-value	$\hat{\beta}_2$	t-test	p-value
07/05/2010	0.187	2.042	0.043	0.674	4.134	0.000	0.380	3.330	0.001	0.650	7.862	0.000
12/31/2010	0.188	1.760	0.081	0.607	4.062	0.000	0.557	5.104	0.000	0.404	7.773	0.000
07/05/2011	0.373	4.039	0.000	0.854	2.590	0.011	0.330	3.163	0.002	0.425	6.960	0.000
01/03/2012	0.219	2.206	0.029	0.722	5.168	0.000	1.031	9.053	0.000	0.643	9.199	0.000
07/03/2012	0.106	1.278	0.204	0.884	6.969	0.000	0.581	5.537	0.000	0.518	7.374	0.000
01/02/2013	0.169	1.762	0.080	0.304	2.777	0.006	0.345	4.163	0.000	0.081	1.120	0.265
07/03/2013	0.244	4.399	0.000	0.371	4.997	0.000	0.493	7.046	0.000	0.453	5.137	0.000
01/02/2014	0.112	2.559	0.012	0.384	5.348	0.000	0.453	6.610	0.000	0.395	5.213	0.000
07/07/2014	0.102	1.301	0.196	0.492	5.709	0.000	0.405	4.426	0.000	0.288	3.458	0.001
01/05/2015	0.340	4.213	0.000	0.814	10.835	0.000	0.721	9.386	0.000	0.492	7.481	0.000
07/06/2015	0.179	2.871	0.005	0.300	4.206	0.000	0.248	3.302	0.001	0.512	7.062	0.000

The results presented in Table 4.1.5 reveal that most of the regression slopes are significant at 5% level of significance. The null hypothesis is therefore rejected at 5% level of significance. The relationship between the third eight pairs is discussed in the next paragraph.

The regression equations for $\Delta\text{LNED}-\Delta\text{LFSR}$, $\Delta\text{LOML}-\Delta\text{LFSR}$, $\Delta\text{LREI}-\Delta\text{LFSR}$, $\Delta\text{LRMH}-\Delta\text{LFSR}$, $\Delta\text{LSBK}-\Delta\text{LFSR}$, $\Delta\text{LSLM}-\Delta\text{LFSR}$, $\Delta\text{LOML}-\Delta\text{LINL}$ and $\Delta\text{LREI}-\Delta\text{LINL}$ are as follows:

$$R_{NED\ i} = \hat{\beta}_1 + \hat{\beta}_2 R_{FSR\ i} + \hat{\mu}_i \quad (4.3.1)$$

$$R_{OML\ i} = \hat{\beta}_1 + \hat{\beta}_2 R_{FSR\ i} + \hat{\mu}_i \quad (4.3.2)$$

$$R_{REI\ i} = \hat{\beta}_1 + \hat{\beta}_2 R_{FSR\ i} + \hat{\mu}_i \quad (4.3.3)$$

$$R_{RMH\ i} = \hat{\beta}_1 + \hat{\beta}_2 R_{FSR\ i} + \hat{\mu}_i \quad (4.3.4)$$

$$R_{SBK\ i} = \hat{\beta}_1 + \hat{\beta}_2 R_{FSR\ i} + \hat{\mu}_i \quad (4.3.5)$$

$$R_{SLM\ i} = \hat{\beta}_1 + \hat{\beta}_2 R_{FSR\ i} + \hat{\mu}_i \quad (4.3.6)$$

$$R_{OML\ i} = \hat{\beta}_1 + \hat{\beta}_2 R_{INL\ i} + \hat{\mu}_i \quad (4.3.7)$$

$$R_{REI\ i} = \hat{\beta}_1 + \hat{\beta}_2 R_{INL\ i} + \hat{\mu}_i \quad (4.3.8)$$

The relationship is presented in Figure 4.1.6.

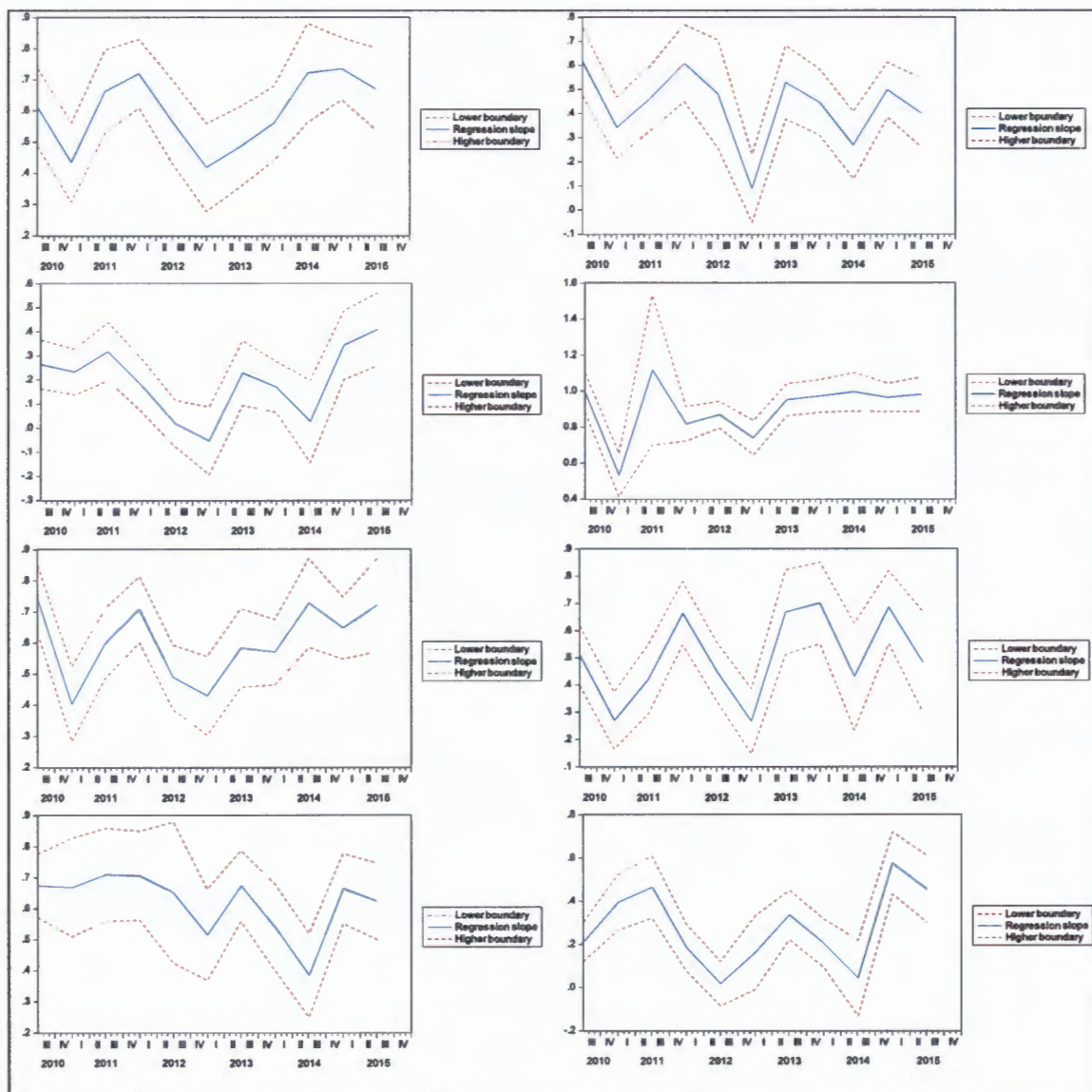


Figure 4.1.6: Beta boundaries for $\Delta\text{LNED}-\Delta\text{LFSR}$, $\Delta\text{LOML}-\Delta\text{LFSR}$, $\Delta\text{LREI}-\Delta\text{LFSR}$, $\Delta\text{LRMH}-\Delta\text{LFSR}$, $\Delta\text{LSBK}-\Delta\text{LFSR}$, $\Delta\text{LSLM}-\Delta\text{LFSR}$, $\Delta\text{LOML}-\Delta\text{LINL}$ and $\Delta\text{LREI}-\Delta\text{LINL}$

Figure 4.1.6 represents the obtained rolling regression slope values for the rolling 125 days. The graph depicts the calculated lower and higher boundaries. The rolling regression slopes of the pair fall between the lower and higher boundaries. The rolling coefficients were further tested for significance and the results are presented in the subsequent table.

Table 4.1.6: Beta boundaries for $\Delta\text{LNED}-\Delta\text{LFSR}$, $\Delta\text{LOML}-\Delta\text{LFSR}$, $\Delta\text{LREI}-\Delta\text{LFSR}$, $\Delta\text{LRMH}-\Delta\text{LFSR}$, $\Delta\text{LSBK}-\Delta\text{LFSR}$, $\Delta\text{LSLM}-\Delta\text{LFSR}$, $\Delta\text{LOML}-\Delta\text{LINL}$ and $\Delta\text{LREI}-\Delta\text{LINL}$

Date	$\Delta\text{LNED}-\Delta\text{LFSR}$			$\Delta\text{LOML}-\Delta\text{LFSR}$			$\Delta\text{LREI}-\Delta\text{LFSR}$			$\Delta\text{LRMH}-\Delta\text{LFSR}$		
	$\hat{\beta}_2$	t-test	p-value	$\hat{\beta}_2$	t-test	p-value	$\hat{\beta}_2$	t-test	p-value	$\hat{\beta}_2$	t-test	p-value
07/05/2010	0.611	9.763	0.000	0.612	8.647	0.000	0.264	5.209	0.000	0.991	18.280	0.000
12/31/2010	0.434	6.833	0.000	0.342	5.389	0.000	0.233	4.889	0.000	0.533	8.950	0.000
07/05/2011	0.662	9.897	0.000	0.464	6.895	0.000	0.318	5.245	0.000	1.116	5.356	0.000
01/03/2012	0.719	13.076	0.000	0.609	7.609	0.000	0.176	3.184	0.002	0.818	17.064	0.000
07/03/2012	0.564	8.412	0.000	0.480	4.276	0.000	0.018	0.372	0.710	0.869	23.370	0.000
01/02/2013	0.418	5.943	0.000	0.089	1.264	0.209	-0.053	-0.756	0.451	0.740	15.425	0.000
07/03/2013	0.484	7.537	0.000	0.529	6.898	0.000	0.228	3.389	0.001	0.952	21.286	0.000
01/02/2014	0.559	9.388	0.000	0.447	6.500	0.000	0.171	3.215	0.002	0.973	21.464	0.000
07/07/2014	0.722	9.072	0.000	0.268	3.852	0.000	0.027	0.316	0.752	0.995	18.581	0.000
01/05/2015	0.735	14.745	0.000	0.498	8.684	0.000	0.344	4.792	0.000	0.964	24.435	0.000
07/06/2015	0.670	10.204	0.000	0.402	5.648	0.000	0.410	5.403	0.000	0.982	20.651	0.000
Date	$\Delta\text{LSBK}-\Delta\text{LFSR}$			$\Delta\text{LSLM}-\Delta\text{LFSR}$			$\Delta\text{LOML}-\Delta\text{LINL}$			$\Delta\text{LREI}-\Delta\text{LINL}$		
	$\hat{\beta}_2$	t-test	p-value	$\hat{\beta}_2$	t-test	p-value	$\hat{\beta}_2$	t-test	p-value	$\hat{\beta}_2$	t-test	p-value
7/05/2010	0.740	13.624	0.000	0.508	9.223	0.000	0.675	13.102	0.000	0.213	4.665	0.000
12/31/2010	0.404	6.723	0.000	0.270	5.163	0.000	0.669	8.413	0.000	0.395	6.122	0.000
7/05/2011	0.600	10.677	0.000	0.422	6.632	0.000	0.710	9.475	0.000	0.466	6.465	0.000
1/03/2012	0.708	13.233	0.000	0.664	11.375	0.000	0.707	9.857	0.000	0.186	3.412	0.001
7/03/2012	0.489	9.396	0.000	0.448	8.264	0.000	0.652	5.729	0.000	0.018	0.358	0.721
1/02/2013	0.430	6.790	0.000	0.267	4.484	0.000	0.516	7.003	0.000	0.161	1.891	0.061
7/03/2013	0.584	9.230	0.000	0.668	8.629	0.000	0.674	11.753	0.000	0.337	5.853	0.000
1/02/2014	0.571	10.809	0.000	0.701	9.332	0.000	0.540	7.724	0.000	0.205	3.651	0.000
7/07/2014	0.728	10.136	0.000	0.430	4.375	0.000	0.385	5.715	0.000	0.044	0.498	0.620
1/05/2015	0.648	12.971	0.000	0.686	10.243	0.000	0.665	11.728	0.000	0.577	8.042	0.000
7/06/2015	0.721	9.674	0.000	0.483	5.188	0.000	0.623	10.089	0.000	0.457	5.872	0.000

The results in Table 4.1.6 present the p-values of the regression slope of the eight pairs. Most of the regression slopes are significant at 5% level of significance. The relationship between the fourth eight pairs is discussed below.

The regression equation for $\Delta\text{LRMH}-\Delta\text{LINL}$, $\Delta\text{LSLM}-\Delta\text{LINL}$, $\Delta\text{LOML}-\Delta\text{LNED}$, $\Delta\text{LREI}-\Delta\text{LNED}$, $\Delta\text{LRMH}-\Delta\text{LNED}$, $\Delta\text{LSBK}-\Delta\text{LNED}$, $\Delta\text{LSLM}-\Delta\text{LNED}$ and $\Delta\text{LREI}-\Delta\text{LOML}$ are as follows:

$$R_{NED\ i} = \hat{\beta}_1 + \hat{\beta}_2 R_{FSR\ i} + \hat{\mu}_i \quad (4.4.1)$$

$$R_{OML\ i} = \hat{\beta}_1 + \hat{\beta}_2 R_{FSR\ i} + \hat{\mu}_i \quad (4.4.2)$$

$$R_{REI\ i} = \hat{\beta}_1 + \hat{\beta}_2 R_{FSR\ i} + \hat{\mu}_i \quad (4.4.3)$$

$$R_{RMH\ i} = \hat{\beta}_1 + \hat{\beta}_2 R_{FSR\ i} + \hat{\mu}_i \quad (4.4.4)$$

$$R_{SBK\ i} = \hat{\beta}_1 + \hat{\beta}_2 R_{FSR\ i} + \hat{\mu}_i \quad (4.4.5)$$

$$R_{SLM\ i} = \hat{\beta}_1 + \hat{\beta}_2 R_{FSR\ i} + \hat{\mu}_i \quad (4.4.6)$$

$$R_{OML\ i} = \hat{\beta}_1 + \hat{\beta}_2 R_{INL\ i} + \hat{\mu}_i \quad (4.4.7)$$

$$R_{REI\ i} = \hat{\beta}_1 + \hat{\beta}_2 R_{INL\ i} + \hat{\mu}_i \quad (4.4.8)$$

The relationship is presented in the following graph.

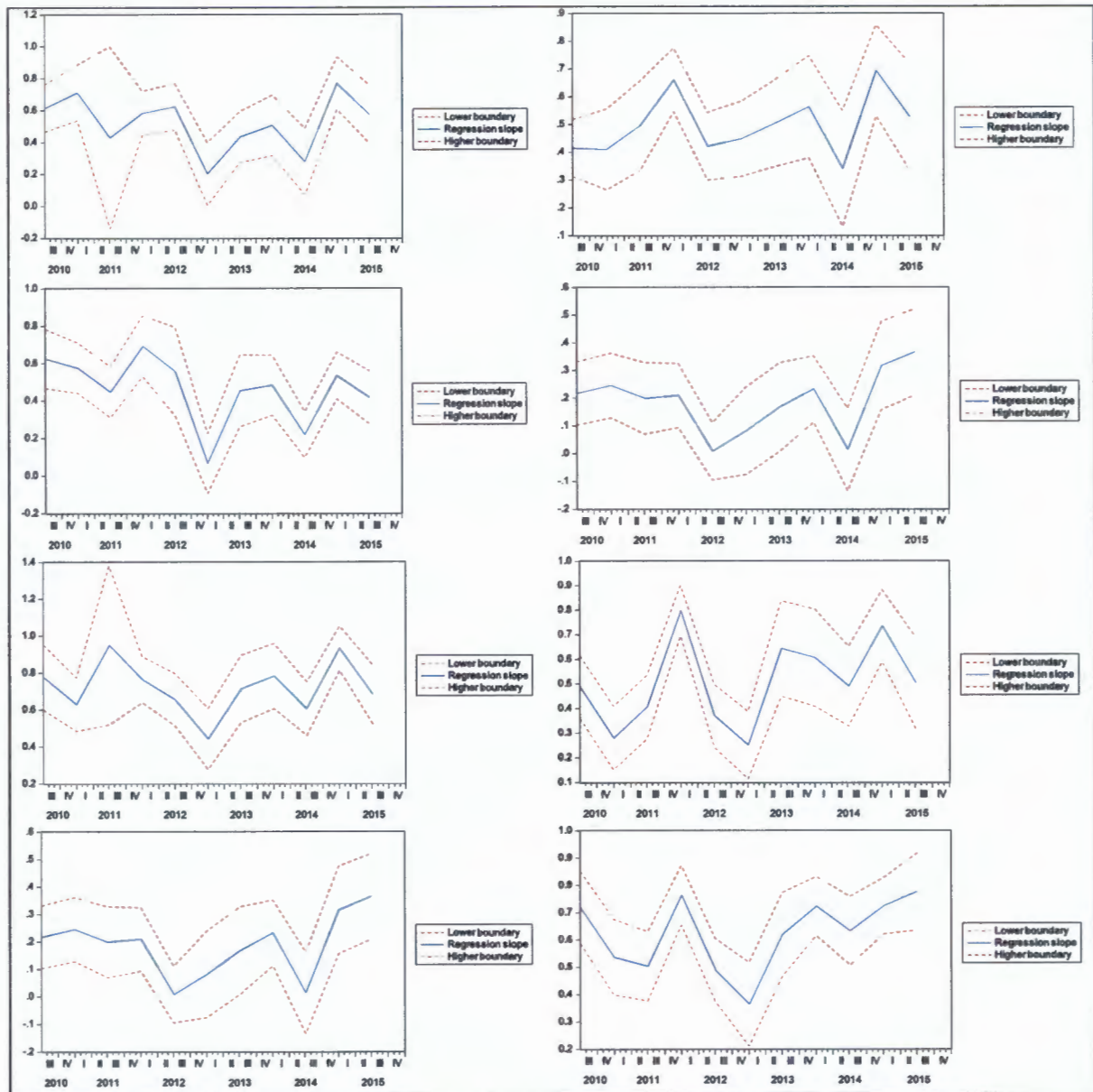


Figure 4.1.7: Beta boundaries for $\Delta\text{LRMH}-\Delta\text{LINL}$, $\Delta\text{LSLM}-\Delta\text{LINL}$, $\Delta\text{LOML}-\Delta\text{LNED}$, $\Delta\text{LSBK}-\Delta\text{LNED}$, $\Delta\text{ALREI}-\Delta\text{LNED}$, $\Delta\text{ALREI}-\Delta\text{LOML}$, $\Delta\text{ALREI}-\Delta\text{LNED}$ and $\Delta\text{ALREI}-\Delta\text{LOML}$

The rolling regression slopes of the eight pairs presented in the above figure falls within the upper and lower 95% confidence limits. The rolling coefficients were tested for significance and the results are presented in Table 4.1.7.

Table 4.1.7: Beta boundaries for $\Delta\text{LRMH}-\Delta\text{LINL}$, $\Delta\text{LSLM}-\Delta\text{LINL}$, $\Delta\text{LOML}-\Delta\text{LNED}$, $\Delta\text{LREI}-\Delta\text{LNED}$, $\Delta\text{LRMH}-\Delta\text{LNED}$, $\Delta\text{LSBK}-\Delta\text{LNED}$, $\Delta\text{LSLM}-\Delta\text{LNED}$ and $\Delta\text{LREI}-\Delta\text{LOML}$

Date	$\Delta\text{LRMH}-\Delta\text{LINL}$			$\Delta\text{LSLM}-\Delta\text{LINL}$			$\Delta\text{LOML}-\Delta\text{LNED}$			$\Delta\text{LREI}-\Delta\text{LNED}$		
	$\hat{\beta}_2$	t-test	p-value	$\hat{\beta}_2$	t-test	p-value	$\hat{\beta}_2$	t-test	p-value	$\hat{\beta}_2$	t-test	p-value
07/05/2010	0.613	8.253	0.000	0.415	8.086	0.000	0.624	7.906	0.000	0.218	3.806	0.000
12/31/2010	0.710	8.142	0.000	0.410	5.637	0.000	0.575	8.577	0.000	0.246	4.203	0.000
07/05/2011	0.429	1.505	0.135	0.495	6.151	0.000	0.446	6.538	0.000	0.199	3.082	0.003
01/03/2012	0.582	8.434	0.000	0.660	11.497	0.000	0.692	8.514	0.000	0.209	3.616	0.000
07/03/2012	0.622	8.544	0.000	0.421	6.895	0.000	0.557	4.711	0.000	0.008	0.164	0.870
01/02/2013	0.201	2.044	0.043	0.447	6.620	0.000	0.066	0.832	0.407	0.082	1.042	0.300
07/03/2013	0.432	5.416	0.000	0.504	6.441	0.000	0.454	4.733	0.000	0.168	2.104	0.037
01/02/2014	0.505	5.337	0.000	0.562	6.122	0.000	0.483	6.005	0.000	0.233	3.878	0.000
07/07/2014	0.278	2.679	0.008	0.340	3.265	0.001	0.220	3.583	0.000	0.015	0.196	0.845
01/05/2015	0.768	9.325	0.000	0.693	8.453	0.000	0.535	8.553	0.000	0.316	3.960	0.000
07/06/2015	0.572	6.300	0.000	0.527	5.480	0.000	0.418	5.881	0.000	0.366	4.689	0.000
Date	$\Delta\text{LRMH}-\Delta\text{LNED}$			$\Delta\text{LSBK}-\Delta\text{LNED}$			$\Delta\text{LSLM}-\Delta\text{LNED}$			$\Delta\text{LREI}-\Delta\text{LOML}$		
	$\hat{\beta}_2$	t-test	p-value	$\hat{\beta}_2$	t-test	p-value	$\hat{\beta}_2$	t-test	p-value	$\hat{\beta}_2$	t-test	p-value
07/05/2010	0.776	8.784	0.000	0.721	10.855	0.000	0.489	7.691	0.000	0.201	3.782	0.000
12/31/2010	0.629	8.702	0.000	0.537	7.778	0.000	0.281	4.361	0.000	0.263	4.263	0.000
07/05/2011	0.950	4.417	0.000	0.504	7.952	0.000	0.407	6.314	0.000	0.364	5.318	0.000
01/03/2012	0.764	12.251	0.000	0.765	13.977	0.000	0.799	15.538	0.000	0.121	2.326	0.022
07/03/2012	0.654	9.241	0.000	0.491	8.522	0.000	0.370	5.801	0.000	-0.020	-0.548	0.585
01/02/2013	0.443	5.356	0.000	0.365	4.780	0.000	0.252	3.684	0.000	0.222	2.560	0.012
07/03/2013	0.713	7.853	0.000	0.621	8.106	0.000	0.644	6.634	0.000	0.387	6.392	0.000
01/02/2014	0.782	8.854	0.000	0.724	13.377	0.000	0.606	6.144	0.000	0.290	5.128	0.000
07/07/2014	0.606	8.338	0.000	0.634	10.072	0.000	0.491	6.048	0.000	0.027	0.257	0.798
01/05/2015	0.933	15.531	0.000	0.725	14.022	0.000	0.736	10.035	0.000	0.574	7.050	0.000
07/06/2015	0.683	8.572	0.000	0.778	11.003	0.000	0.504	5.426	0.000	0.519	6.313	0.000

Table 4.1.7 depicts the t-tests and p-values of the rolling coefficients. Most of the estimated slope coefficients were found to be significantly different from zero at a 95% confidence level. The results also revealed that most of the p-values of the coefficients are all less than 5% level of significance. The relationship between the last seven pairs is discussed in the next paragraph.

The regression equation for $\Delta\text{LRMH}-\Delta\text{LOML}$, $\Delta\text{LSLM}-\Delta\text{LOML}$, $\Delta\text{LRMH}-\Delta\text{LREI}$, $\Delta\text{LSLM}-\Delta\text{LREI}$, $\Delta\text{LSBK}-\Delta\text{LRMH}$, $\Delta\text{LSLM}-\Delta\text{LRMH}$ and $\Delta\text{LSLM}-\Delta\text{LSBK}$ are as follows:

$$R_{RMH\ i} = \hat{\beta}_1 + \hat{\beta}_2 R_{OML\ i} + \hat{\mu}_i \quad (4.5.1)$$

$$R_{SLM\ i} = \hat{\beta}_1 + \hat{\beta}_2 R_{OML\ i} + \hat{\mu}_i \quad (4.5.2)$$

$$R_{RMH\ i} = \hat{\beta}_1 + \hat{\beta}_2 R_{REI\ i} + \hat{\mu}_i \quad (4.5.3)$$

$$R_{SLM\ i} = \hat{\beta}_1 + \hat{\beta}_2 R_{REI\ i} + \hat{\mu}_i \quad (4.5.4)$$

$$R_{SBK\ i} = \hat{\beta}_1 + \hat{\beta}_2 R_{RMH\ i} + \hat{\mu}_i \quad (4.5.5)$$

$$R_{SLM\ i} = \hat{\beta}_1 + \hat{\beta}_2 R_{RMH\ i} + \hat{\mu}_i \quad (4.5.6)$$

$$R_{SLM\ i} = \hat{\beta}_1 + \hat{\beta}_2 R_{SBK\ i} + \hat{\mu}_i \quad (4.5.7)$$

The relationship is presented in Figure 4.1.8.

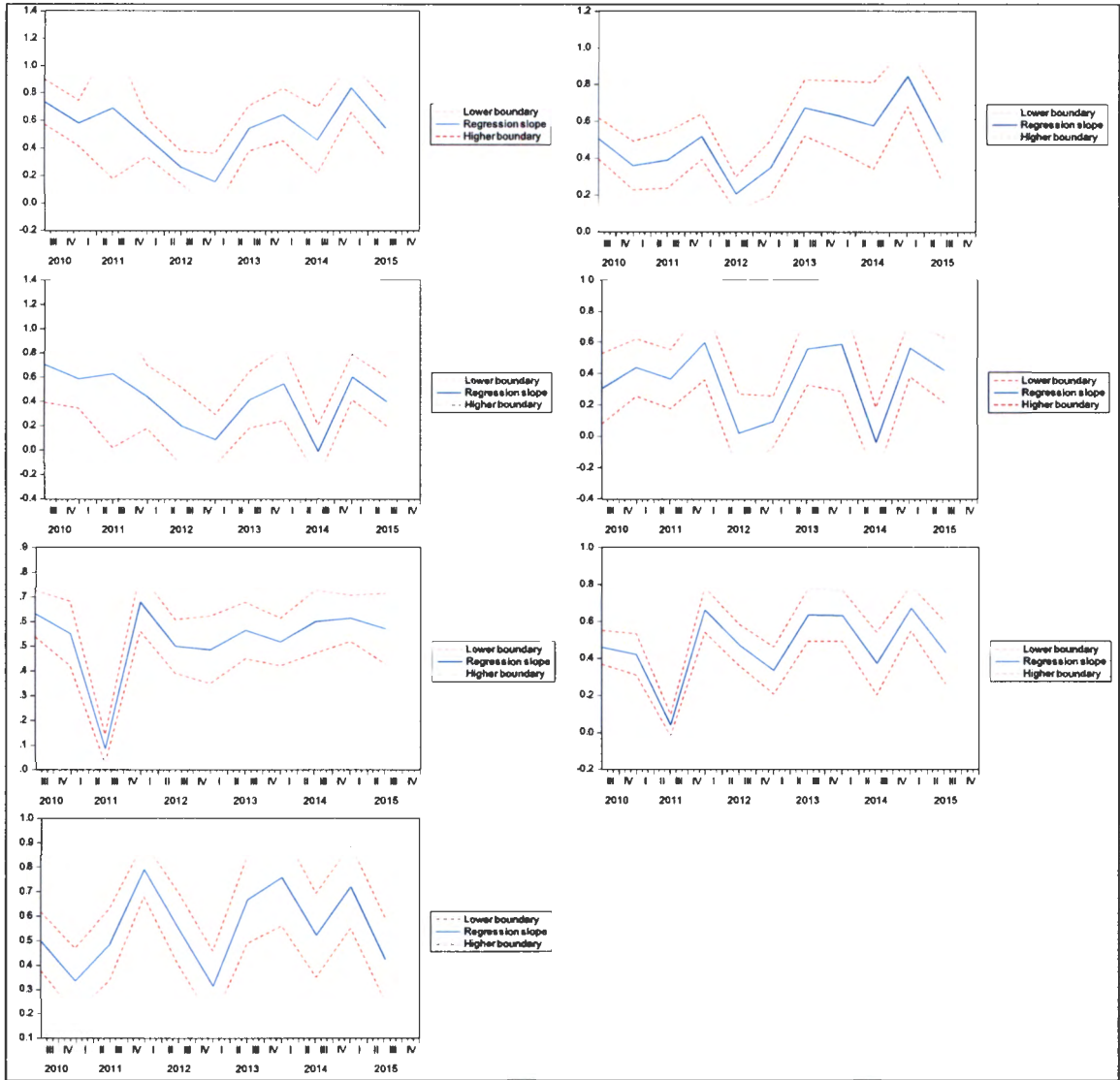


Figure 4.1.8: Beta boundaries for $\Delta\text{LRMH}-\Delta\text{LOML}$, $\Delta\text{LSLM}-\Delta\text{LOML}$, $\Delta\text{LRMH}-\Delta\text{LREI}$, $\Delta\text{LSLM}-\Delta\text{LREI}$, $\Delta\text{LSBK}-\Delta\text{LRMH}$, $\Delta\text{LSLM}-\Delta\text{LRMH}$ and $\Delta\text{LSLM}-\Delta\text{LSBK}$

The graphical presentation presented in the above figure revealed that the rolling regression slopes of the seven pairs falls within the estimated 95% confidence interval. The rolling coefficients were further tested for significance and the results are presented in Table 4.1.8.

Table 4.1.8: Beta boundaries for $\Delta\text{LRMH}-\Delta\text{LOML}$, $\Delta\text{LSLM}-\Delta\text{LOML}$, $\Delta\text{LRMH}-\Delta\text{LREI}$, $\Delta\text{LSLM}-\Delta\text{LREI}$, $\Delta\text{LSBK}-\Delta\text{LRMH}$, $\Delta\text{LSLM}-\Delta\text{LRMH}$ and $\Delta\text{LSLM}-\Delta\text{LSBK}$

Date	$\Delta\text{LRMH}-\Delta\text{LOML}$			$\Delta\text{LSLM}-\Delta\text{LOML}$			$\Delta\text{LRMH}-\Delta\text{LREI}$			$\Delta\text{LSLM}-\Delta\text{LREI}$		
	$\hat{\beta}_2$	t-test	p-value	$\hat{\beta}_2$	t-test	p-value	$\hat{\beta}_2$	t-test	p-value	$\hat{\beta}_2$	t-test	p-value
07/05/2010	0.734	9.061	0.000	0.506	9.140	0.000	0.704	4.542	0.000	0.304	2.731	0.007
12/31/2010	0.579	7.061	0.000	0.360	5.476	0.000	0.586	4.864	0.000	0.437	4.796	0.000
07/05/2011	0.687	2.703	0.008	0.389	5.109	0.000	0.627	2.065	0.041	0.364	3.906	0.000
01/03/2012	0.476	6.912	0.000	0.518	8.410	0.000	0.440	3.379	0.001	0.596	5.056	0.000
07/03/2012	0.261	4.375	0.000	0.207	4.433	0.000	0.199	1.256	0.212	0.021	0.168	0.867
01/02/2013	0.156	1.526	0.130	0.347	4.651	0.000	0.086	0.829	0.409	0.095	1.181	0.240
07/03/2013	0.542	6.595	0.000	0.672	8.827	0.000	0.414	3.554	0.001	0.555	4.864	0.000
01/02/2014	0.641	6.790	0.000	0.629	6.639	0.000	0.545	3.602	0.000	0.585	3.910	0.000
07/07/2014	0.456	3.826	0.000	0.575	4.900	0.000	-0.012	-0.108	0.914	-0.037	-0.336	0.737
01/05/2015	0.836	9.309	0.000	0.844	10.217	0.000	0.601	6.373	0.000	0.562	6.161	0.000
07/06/2015	0.543	5.389	0.000	0.487	4.575	0.000	0.399	3.990	0.000	0.420	4.100	0.000
Date	$\Delta\text{LSBK}-\Delta\text{LRMH}$			$\Delta\text{LSLM}-\Delta\text{LRMH}$			$\Delta\text{LSLM}-\Delta\text{LSBK}$					
	$\hat{\beta}_2$	t-test	p-value	$\hat{\beta}_2$	t-test	p-value	$\hat{\beta}_2$	t-test	p-value			
07/05/2010	0.633	13.371	0.000	0.461	10.090	0.000	0.499	8.340	0.000			
12/31/2010	0.552	8.408	0.000	0.422	7.550	0.000	0.337	5.023	0.000			
07/05/2011	0.086	2.956	0.004	0.043	1.517	0.132	0.486	6.636	0.000			
01/03/2012	0.680	11.508	0.000	0.664	10.872	0.000	0.790	14.237	0.000			
07/03/2012	0.500	9.144	0.000	0.473	8.471	0.000	0.550	7.445	0.000			
01/02/2013	0.485	7.091	0.000	0.336	5.297	0.000	0.315	4.370	0.000			
07/03/2013	0.564	9.851	0.000	0.636	8.946	0.000	0.666	7.535	0.000			
01/02/2014	0.517	10.680	0.000	0.632	9.143	0.000	0.757	7.717	0.000			
07/07/2014	0.600	9.383	0.000	0.374	4.422	0.000	0.524	6.129	0.000			
01/05/2015	0.613	13.072	0.000	0.672	11.008	0.000	0.720	8.524	0.000			
07/06/2015	0.570	7.919	0.000	0.431	5.164	0.000	0.424	4.991	0.000			

Table 4.1.8 depicts the t-tests and p-values of the rolling coefficients. The results revealed that most of the p-values of the regression slopes are significant at 5% level of significance.

The OLS approach only revealed that variables are related but it does not tell whether or not they are related in a long run. Therefore, the next step is to test whether or not there is a long run relationship amongst the paired indexes using the Engle and Granger (1987) cointegration

technique. The pairs were tested for long run relationship and the results are presented in the following section.

4.2.5 Verification of cointegration

The Engle and Granger (EG) cointegration test was applied to the thirty nine paired indexes. The cointegration methodology is used to check if the preselected stock pairs share any long term equilibrium relationship. The results are presented in Table 4.1.9.

Table 4.1.9: Engle and Granger results for 39 paired indexes in financial sector

<i>Pairs</i>	<i>OLS Results</i>		<i>Residuals from OLS</i>	
	β_1 Coefficient	<i>t</i> -statistic	<i>ADF (t-statistic)</i>	<i>p-value</i>
LDSY-LBVT	1.470	224.285	-4.243	0.001*
LFSR-LBVT	1.110	142.898	-2.781	0.061***
LINL-LBVT	0.845	68.1725	-1.732	0.415
LNED-LBVT	0.695	130.590	-2.414	0.138
LOML-LBVT	1.151	205.853	-4.473	0.000*
LREI-LBVT	1.018	182.833	-3.870	0.002***
LRMH-LBVT	0.880	80.934	-2.281	0.178
LSLM-LBVT	1.203	208.520	-4.178	0.001*
LFSR-LDSY	0.749	158.916	-2.712	0.072***
LINL-LDSY	0.582	76.845	-1.950	0.309
LNED-LDSY	0.459	110.451	-1.983	0.294
LOML-LDSY	0.769	188.823	-3.864	0.002***
LREI-LDSY	0.683	186.152	-3.480	0.009***
LRMH-LDSY	0.598	87.921	-2.417	0.137
LSLM-LDSY	0.798	163.443	-2.335	0.161
LINL-LFSR	0.768	84.400	-1.634	0.465
LNED-LFSR	0.611	153.090	-2.773	0.062***
LOML-LFSR	0.978	132.368	-2.777	0.062***
LREI-LFSR	0.872	141.865	-2.868	0.049***
LRMH-LFSR	0.807	120.582	-2.402	0.141
LSBK-LFSR	0.387	90.628	-2.490	0.118
LSLM-LFSR	1.040	180.954	-3.683	0.005***
LOML-LINL	1.061	71.826	-1.670	0.447
LREI-LINL	0.969	82.424	-1.975	0.298
LRMH-LINL	0.936	99.880	-2.744	0.067***
LSLM-LINL	1.106	71.288	-1.824	0.369
LOML-LNED	1.542	123.767	-2.109	0.241

<i>Pairs</i>	<i>OLS Results</i>		<i>Residuals from OLS</i>	
	β_1 Coefficient	<i>t</i> -statistic	<i>ADF (t-statistic)</i>	<i>p-value</i>
LREI-LNED	1.351	107.924	-1.623	0.471
LRMH-LNED	1.198	75.843	-1.823	0.370
LSBK-LNED	0.591	73.889	-2.367	0.151
LSLM-LNED	1.647	173.067	-3.417	0.011*
LREI-LOML	0.863	159.499	-2.988	0.036**
LRMH-LOML	0.752	81.502	-2.405	0.141
LSLM-LOML	1.014	158.058	-3.092	0.027**
LRMH-LREI	0.843	79.375	-2.381	0.147
LSLM-LREI	1.139	149.987	-2.995	0.036**
LSBK-LRMH	0.461	96.219	-2.745	0.067***
LSLM-LRMH	1.147	87.410	-2.405	0.140
LSLM-LSBK	2.212	70.082	-1.791	0.385

Note: *, **, and *** indicates the MacKinnon critical values at 1%, 5% and 10% levels are -3.435, -2.863, and -2.568 respectively. Lag length was selected by SIC.

Table 4.1.9 presents the results of the EG cointegration between the preselected pairs. The results revealed that eighteen pairs are cointegrated since their critical values are significant at 1%, 5% and 10% levels of significance. According to Asteriou and Hall (2011), if variables are cointegrated, their long run relationship is not spurious. Those eighteen pairs are used in subsequent analyses to implement pairs trading strategy.

4.2.6 Error Correction Model

The Granger representation states that when two time series are cointegrated, a vector autoregressive (VAR) model on the differences would be mis-specified (Granger, 1986). Therefore, the ECM is a dynamic model that reconciles the long run equilibrium with disequilibrium behaviour in the short run. The ECM results are summarised in the Table 4.1.10.

Table 4.1.10: ECM results for the cointegrated pairs of the financial sector

<i>Cointegrated pairs</i>	<i>Stock price</i>	<i>Speed of adjustment</i>		<i>Cointegrated pairs</i>	<i>Stock price</i>	<i>Speed of adjustment</i>	
		γ_1	γ_2			γ_1	γ_2
LDSY-LBVT	LDSY	-0.018**		LOML-LFSR	LOML	-0.009**	
	LBVT		0.018**		LFSR		0.008**
LFSR-LBVT	LFSR	-0.010**		LREI-LFSR	LREI	-0.007**	
	LBVT		0.010**		LFSR		0.012**
LOML-LBVT	LOML	-0.023**		LSLM-LFSR	LSLM	-0.011**	
	LBVT		0.020**		LFSR		0.016**
LREI-LBVT	LREI	-0.014**		LRMH-LINL	LRMH	-0.010**	
	LBVT		0.020**		LINL		0.011**
LSLM-LBVT	LSLM	-0.019**		LSLM-LNED	LSLM	-0.013**	
	LBVT		0.016**		LNED		0.015**
LFSR-LDSY	LFSR	-0.012***		LREI-LOML	LREI	-0.010**	
	LDSY		0.011**		LOML		0.013**
LOML-LDSY	LOML	-0.018**		LSLM-LOML	LSLM	-0.008**	
	LDSY		0.012**		LOML		0.012**
LREI-LDSY	LREI	-0.014**		LSLM-LREI	LSLM	-0.010**	
	LDSY		0.016**		LREI		0.008**
LNED-LFSR	LNED	-0.018**		LSBK-LRMH	LSBK	-0.019**	
	LFSR		0.022**		LRMH		0.021**

Note: *, **, *** represents the significant level at 10%, 5% and 1%, respectively

The results in Table 4.1.10 revealed that the disequilibrium terms (γ_1 and γ_2) of the cointegrated pair are significantly different from zero, which confirms the earlier results of EG cointegration test. The EG cointegration results confirmed that the eighteen variables are cointegrated, therefore, the short term disequilibrium term was used to correct the system over the long term equilibrium term. The size and sign of γ_1 and γ_2 are important as an indication of mean reversion property towards long run disequilibrium. For this reason, the focus is on the correct sign (-) and size. The error correction term with a correct sign are all significant at 5% level of significance.

The next step is to compute the model diagnostics. Although it is a practice in econometrics time series to do diagnostics after presenting short run results, this is not normally done in the application of statistical arbitrage since arbitrage deals with financial data. Most studies using financial data have not concerned themselves with diagnostic tests due to the nature of the

financial data. If the data is not consistent and stable, as is the common case with financial data, this step is normally not computed. According to Balarezo (2010), financial data also has the characteristic of residuals not being normally distributed. They are either negatively skewed or positively skewed. This was revealed by the results in Table 4.1.11.1.

4.2.7 Model diagnostics results

The models are tested for mis-specification by checking whether the errors in the model are normally distributed or they show any ARCH effect or serial correlation. The residuals are said to be uncorrelated if the p-value of the test statistic is greater than 5% level of significance, there is a presence of heteroscedasticity if the p-value of LM test is less than 5% level of significance and the residuals are normally distributed if the p-value of JB test is more than 5% level of significance. The diagnostic tests are performed on the eighteen cointegrated pairs and the results are summarised in the Table 4.1.11.1, Table 4.1.11.2 and Table 4.1.11.3.

Table 4.1.11.1: Diagnostic test results for LDSY-LBVT, LFSR-LBVT, LOML-LBVT, LREI-LBVT, LSLM-LBVT, LFSR-LDSY, LOML-LDSY and LREI-LDSY

<i>Test</i>	<i>LDSY-LBVT</i>		<i>LFSR-LBVT</i>		<i>LOML-LBVT</i>		<i>LREI-LBVT</i>	
	Test statistic	p-value	Test statistic	p-value	Test statistic	p-value	Test statistic	p-value
Breusch-Godfrey (BG)	4.712	0.095	24.154	0.063	8.324	0.080	67.151	0.051
Lagrange Multiplier (LM)	13.582	0.001	2.297	0.317	1.501	0.472	10.700	0.005
Jarque-Bera (JB)	4690.173	0.000	4617.604	0.000	20005	0.000	36.568	0.000
Skewness	-0.565		-0.796		-0.872		0.163	
Kurtosis	11.592		11.450		20.812		3.692	
ARCH	13.6311	0.000					28.819	0.000
Test	LSLM-LBVT		LFSR-LDSY		LOML-LDSY		LREI-LDSY	
	Test statistic	p-value	Test statistic	p-value	Test statistic	p-value	Test statistic	p-value
Breusch-Godfrey (BG)	1.698	0.428	185.468	0.054	3.100	0.212	81.844	0.056
Lagrange Multiplier (LM)	12.865	0.002	10.567	0.005	2.100	0.350	16.522	0.000
Jarque-Bera (JB)	400.601	0.000	703.545	0.000	19772.10	0.000	42.709	0.000
Skewness	-0.255		-0.398		-0.981		0.194	
Kurtosis	5.481		6.261		20.684		3.730	
ARCH	83.670	0.000	68.867	0.000			30.806	0.000

The results in Table 4.1.11.1 revealed no presence of serial correlation in model residuals of the eight pairs since the p-values of BG test are more than 5% level of significance. There is also no presence of heteroscedasticity in the three pairs (LFSR-LBVT, LOML-LBVT and LOML-LDSY) since the p-values of the LM test statistic are greater than 5% level of significance. The LM test and ARCH LM test revealed a presence of heteroscedasticity in LDSY-LBVT, LREI-LBVT, LSLM-LBVT, LFSR-LDSY and LREI-LDSY pairs. The p-values of JB test are all 0.000. This implies that the data is not normally distributed. The value of the kurtosis test indicates that all the variables have leptokurtosis effect. The results are not surprising since the study models financial data. According to Cont (2001), normal distribution is not possible in most of the financial time series data.

Table 4.1.11.2: Diagnostic test results for LNED-LFSR, LOML-LFSR, LREI-LFSR, LSLM-LFSR, LRMH-LINL, LSLM-LNED, LREI-LOML and LSLM-LOML

<i>Test</i>	<i>LNED-LFSR</i>		<i>LOML-LFSR</i>		<i>LREI-LFSR</i>		<i>LSLM-LFSR</i>	
	Test statistic	p-value	Test statistic	p-value	Test statistic	p-value	Test statistic	p-value
Breusch-Godfrey (BG)	146.309	0.052	4.615	0.100	63.5021	0.055	4.073	0.131
Lagrange Multiplier (LM)	7.6192	0.022	2.177	0.337	23.102	0.000	18.939	0.000
Jarque-Bera (JB)	158.244	0.000	20711.25	0.000	51.477	0.000	196.076	0.000
Skewness	0.082		-0.957		0.177		0.122	
Kurtosis	4.583		21.109		3.836		4.755	
ARCH	9.947458	0.002			41.725	0.000	3.821	0.051
Test	LRMH-LINL		LSLM-LNED		LREI-LOML		LSLM-LOML	
	Test statistic	p-value	Test statistic	p-value	Test statistic	p-value	Test statistic	p-value
Breusch-Godfrey (BG)	47.467	0.062	4.071	0.131	76.381	0.053	4.461	0.108
Lagrange Multiplier (LM)	1.730	0.421	31.799	0.000	21.790	0.000	58.759	0.000
Jarque-Bera (JB)	1160029.	0.000	96.463	0.000	57.591	0.000	627.550	0.000
Skewness	-6.442		0.048		0.209		0.131	
Kurtosis	138.672		4.200		3.865		6.159	
ARCH			19.051	0.000	43.102	0.000	25.746	0.000

The p-values of BG test revealed serial correlation in the residual models presented in Table 4.1.11.2 since their p-values are significant at the 5% level of significance. The LM p-values

of LOML-LFSR and LRMH-LINL pair are significant, therefore the null hypothesis is accepted. This implies that there is no heteroscedasticity. The ARCH LM test was employed to the series with the presence of heteroscedasticity and the results provided evidence of autoregressive conditional heteroscedasticity in LNED-LFSR, LREI-LFSR, LSLM-LNED, LREI-LOML and LSLM-LOML since their p-values are less than 5 % level of significance. The ARCH LM test further revealed no evidence of autoregressive conditional heteroscedasticity in LSLM-LFSR pair. All the pairs are not normally distributed since the p-values of the JB test statistic are less than 5% level of significance. All the eight variables are flat tailed since their values of kurtosis are greater than 3. Yonis (2011) indicated that the financial series has features of fat tails to the normal distribution.

Table 4.1.11.3: Diagnostic test results for LSLM-LREI and LSBK-LRMH

<i>Test</i>	<i>LSLM-LREI</i>		<i>LSBK-LRMH</i>	
	Test statistic	p-value	Test statistic	p-value
Breusch-Godfrey (BG)	1.552	0.460	168.671	0.056
Lagrange Multiplier (LM)	52.672	0.000	375.190	0.000
Jarque-Bera (JB)	1413.462	0.000	50705.670	0.000
Skewness	-0.470		1.636	
Kurtosis	7.665		31.304	
ARCH	189.955	0.000	0.001	0.972

The results in Table 4.1.11.3 reveal that there is no evidence of serial correlation in LSLM-LREI and LSBK-LRMH pair since the p-value of 0.460 and 0.056 are greater than 5% level of significance. The LM p-value of LSLM-LREI pair of 0.000 is less than 5% level of significance, therefore the null hypothesis is rejected. This means that there is a presence of heteroscedasticity. The ARCH LM test also reveals there is autoregressive conditional heteroscedasticity in LSLM-LREI pair. The ARCH LM test of LSBK-LRMH pair revealed no evidence of autoregressive conditional heteroscedasticity since the p-value is greater than 5% level of significance. The p-value of JB test statistic is 0.000 for both LSLM-LREI and LSBK-LRMH pair. The residuals have a leptokurtosis effect since the kurtosis value of 7.665 and 31.304 are greater than the threshold of 3. The results are in line with the views by Yonis (2011). Therefore, it is concluded that the residuals for both the pairs are not normally distributed.

4.2.8 Granger Causality results

The Granger causality technique tests whether or not adding lags of one variable have a predictive power for another variable. The causal relationship between paired stock prices may be unidirectional, bidirectional or independent. The Granger causality results are presented in Table 4.1.12.1 and Table 4.1.12.2

Table 4.1.12.1: Causality results for LBVT-LDSY, LBVT-LFSR, LBVT-LOML, LBVT-LREI, LBVT-LSLM, LDSY-LFSR, LDSY-LOML, LDSY-LREI and LFSR-LNED

Null Hypothesis:	Obs	F-Statistic	Prob.
LBVT does not Granger Cause LDSY	1498	3.098	0.045
LDSY does not Granger Cause LBVT		4.504	0.011
LBVT does not Granger Cause LFSR	1498	0.670	0.512
LFSR does not Granger Cause LBVT		3.857	0.021
LBVT does not Granger Cause LOML	1498	3.768	0.023
LOML does not Granger Cause LBVT		2.752	0.064
LBVT does not Granger Cause LREI	1498	3.267	0.038
LREI does not Granger Cause LBVT		3.770	0.023
LBVT does not Granger Cause LSLM	1498	2.076	0.126
LSLM does not Granger Cause LBVT		1.058	0.347
LDSY does not Granger Cause LFSR	1498	1.501	0.223
LFSR does not Granger Cause LDSY		2.544	0.079
LDSY does not Granger Cause LOML	1498	3.724	0.024
LOML does not Granger Cause LDSY		0.603	0.547
LDSY does not Granger Cause LREI	1498	3.030	0.049
LREI does not Granger Cause LDSY		2.366	0.094
LFSR does not Granger Cause LNED	1498	0.891	0.410
LNED does not Granger Cause LFSR		3.052	0.048

The results in Table 4.1.12.1 revealed that there is a bidirectional causality between LBVT-LDSY and LBVT-LREI since the p-values of the null and alternative hypotheses are less than 5% level of significance. There is a unidirectional causality running from LFSR to LBVT, from LBVT to LOML, from LDSY to LOML, from LDSY to LREI and from LNED to LFSR. There is no evidence of long run causality between LBVT and LSLM, and LDSY and LREI.

Table 4.1.12.2: Causality results for LFSR-LOML, LFSR-LREI, LFSR-LSLM, LINL-LRMH, LNED-LSLM, LOML-LREI, LOML-LSLM, LREI-LSLM and LRMH-LSBK

Null Hypothesis:	Obs	F-Statistic	Prob.
LFSR does not Granger Cause LOML	1498	2.344	0.096
LOML does not Granger Cause LFSR		0.286	0.752
LFSR does not Granger Cause LREI	1498	0.791	0.454
LREI does not Granger Cause LFSR		3.209	0.041
LFSR does not Granger Cause LSLM	1498	0.818	0.442
LSLM does not Granger Cause LFSR		3.220	0.040
LINL does not Granger Cause LRMH	1498	0.121	0.886
LRMH does not Granger Cause LINL		3.967	0.019
LNED does not Granger Cause LSLM	1498	0.441	0.644
LSLM does not Granger Cause LNED		2.572	0.077
LOML does not Granger Cause LREI	1498	2.608	0.074
LREI does not Granger Cause LOML		1.742	0.176
LOML does not Granger Cause LSLM	1498	0.079	0.924
LSLM does not Granger Cause LOML		2.971	0.052
LREI does not Granger Cause LSLM	1498	2.287	0.102
LSLM does not Granger Cause LREI		1.609	0.200
LRMH does not Granger Cause LSBK	1498	2.187	0.113
LSBK does not Granger Cause LRMH		0.883	0.414

The results in Table 4.1.12.2 revealed no evidence of long run causality between LFSR and LOML. There is a unidirectional causality running from LREI to LFSR, from LSLM to LFSR and also from LRMH to LINL. There is no causality between LNED and LSLM, and between LOML and LREI. The results further revealed no long run causality between LOML and LSLM; LREI and LSLM; and LRMH and LSBK pairs.

4.2.9 Implementing pairs trading strategy

This section gives a detailed analysis of the empirical results of the cointegrated pairs. The historic movement of the pairs are presented in Figure 4.1.9, Figure 4.1.10 and Figure 4.1.11. Before implementing pairs trading strategy, it is vital to find out that if the stock returns have similar historic patterns or not. The idea of pairs trading is to buy the relatively undervalued stock and sell the relatively overvalued stock, then close trade when the ratio goes back to the mean ration (line 0). The trading rule used is 2 standard deviation.

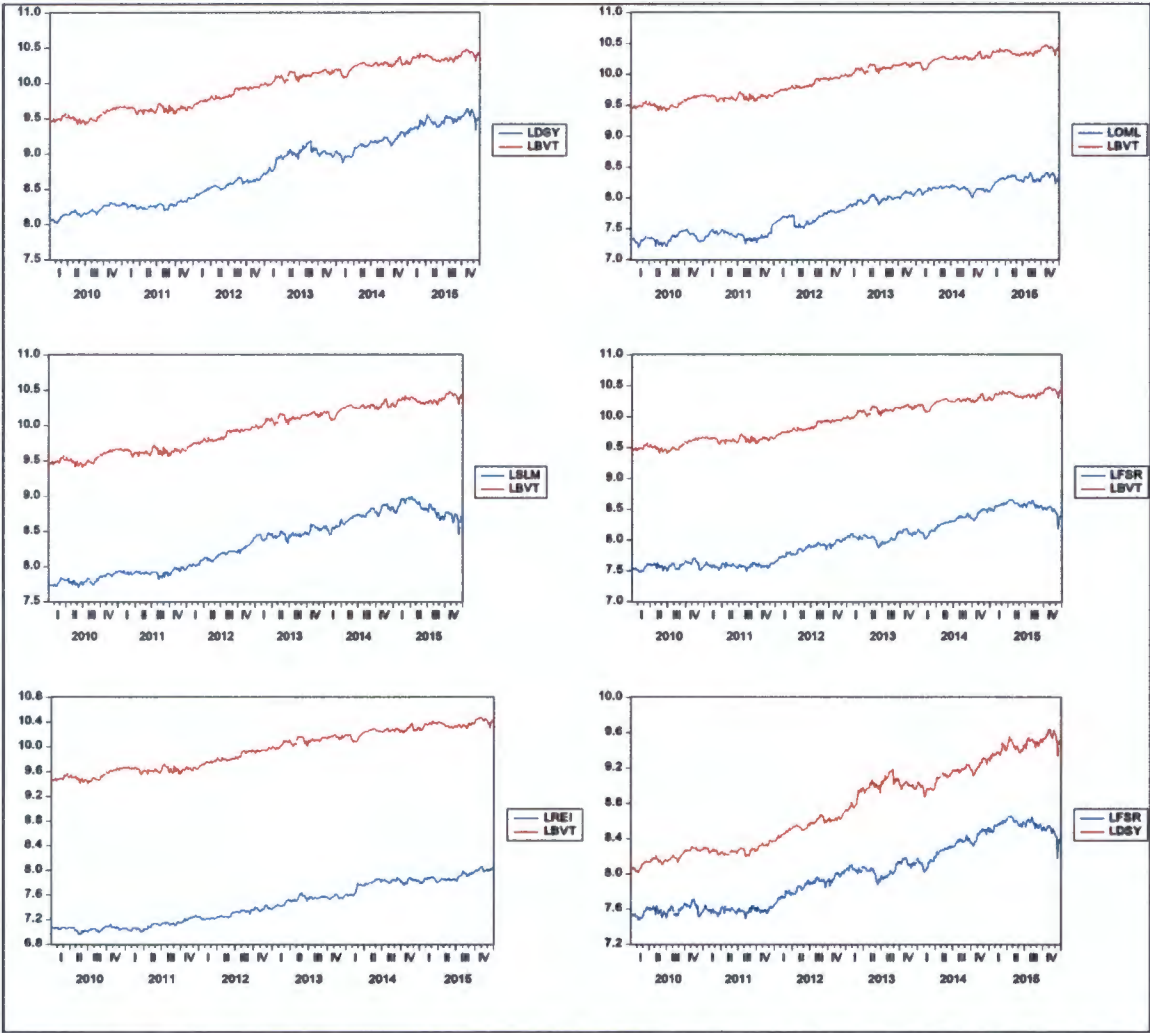
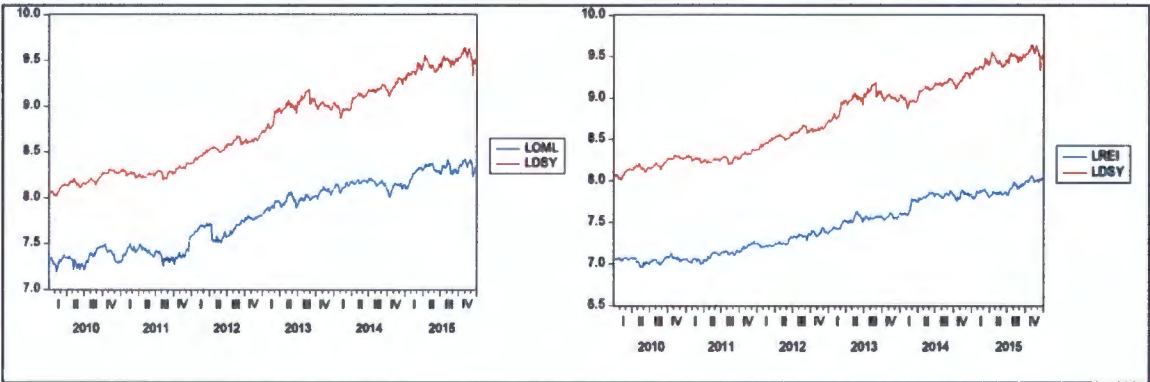


Figure 4.1.9: Historical movement of LDSY-LBVT, LOML-LBVT, LSLM-LBVT, LFSR-LBVT, LREI-LBVT and LFSR-LDSY pairs

Figure 4.1.9 depicts that there is a close relationship between all the stock pairs. The stock prices presented in Figure 4.1.9 tend to move together throughout the sample period. Therefore, pairs trading strategy can be implemented to the above stocks.



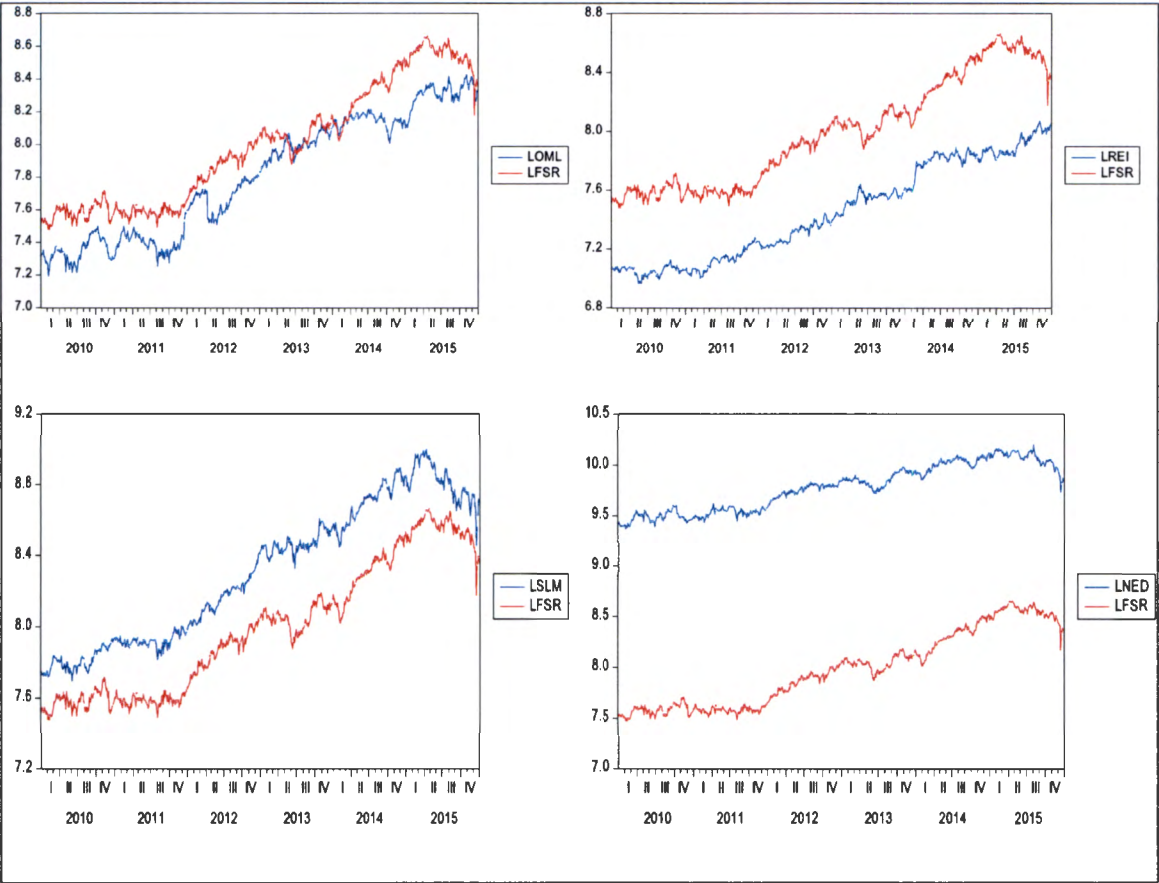
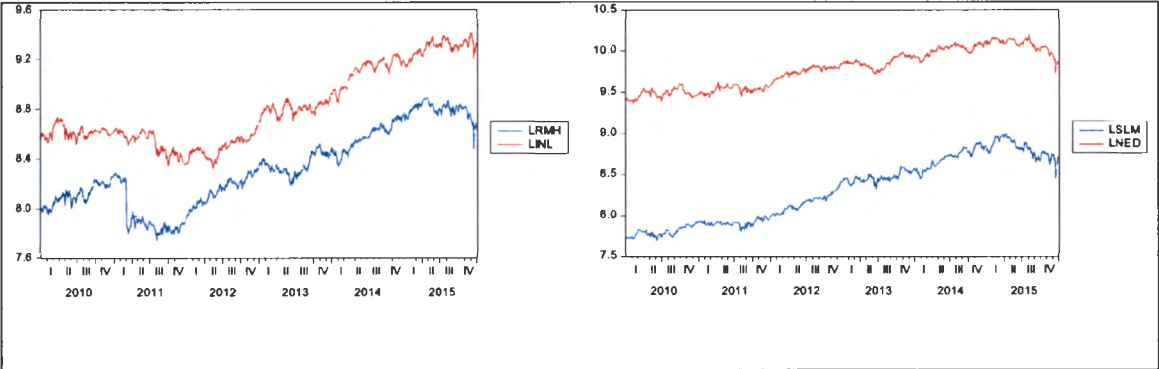


Figure 4.1.10: Historical movement of LOML-LDSY, LREI-LDSY, LOML-LFSR, LREI-LFSR, LSLM-LFSR and LNED-LFSR pairs

The six stock pairs presented in Figure 4.1.10 depict that each pair seems to have a similar pattern throughout the sample period. The LOML-LFSR pair is very closely related compared to other pairs. Therefore, there is a chance for arbitrage operations to take place for LOML-LFSR pair.



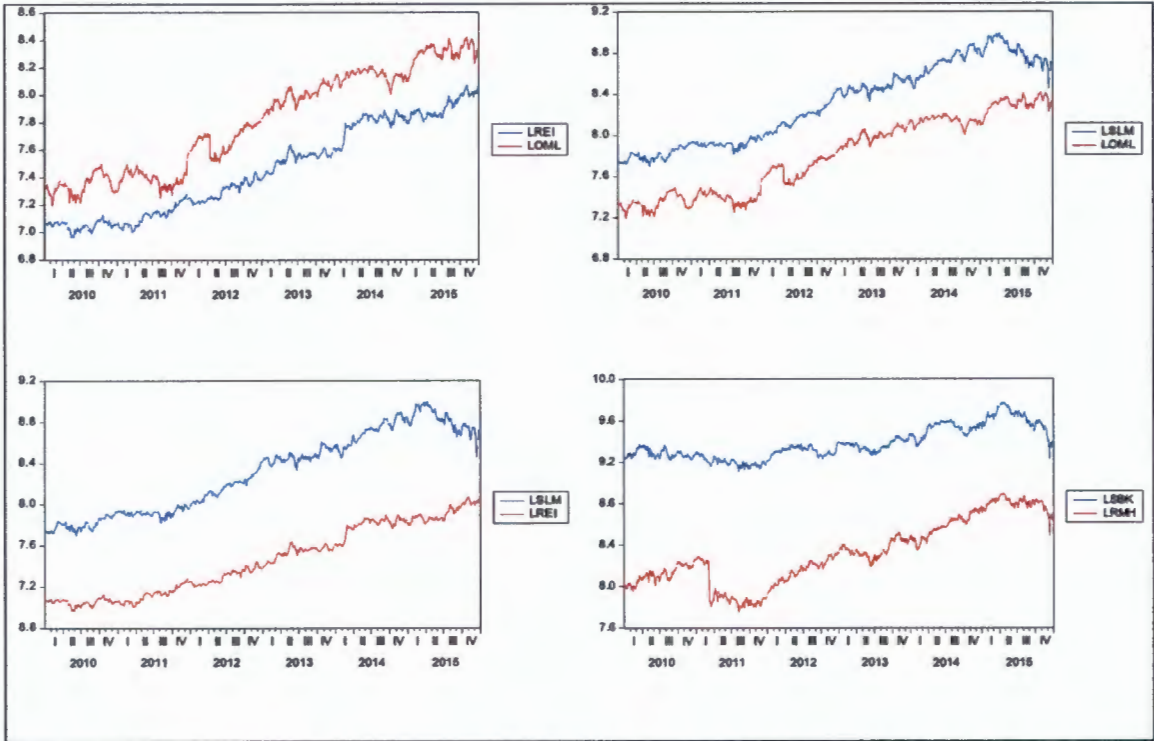
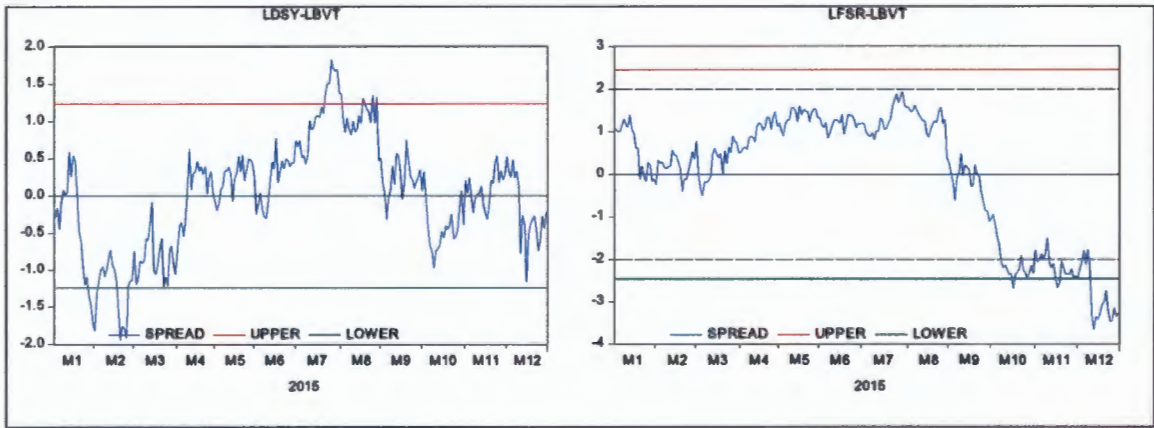


Figure 4.1.11: Historical movement of LRMH-LINL, LSLM-LNED, LREI-LOML, LSLM-LOML, LSLM-LREI AND LSBK-LRMH pairs

The stock prices presented in the pictorial representation in Figure 4.1.11 depicts the historical movement of each of the six stock pairs in the long run. The results in the graphs are in relation to the EG cointegration results. The next step is to plot the residual spread series of the cointegrated pairs to test whether trade can be opened or not.



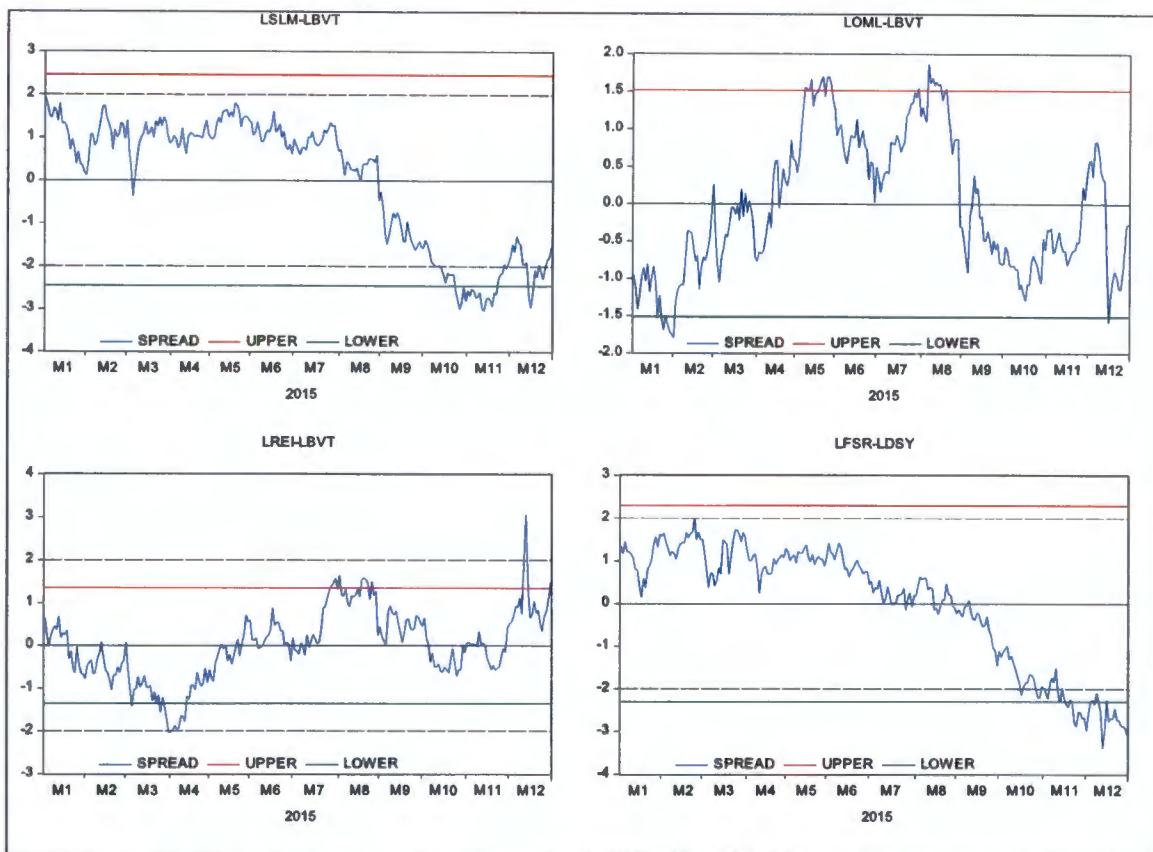


Figure 4.1.12: Spread of LDSY-LBVT, LOML-LBVT, LSLM-LBVT, LFSR-LBVT, LREI-LBVT and LFSR-LDSY pairs with trading thresholds

The pairs trading was implemented to the data from January 2015 to December 2015. Figure 4.1.12 depicts the ratio of LDSY to LBVT, LOML to LBVT, LFSR to LBVT, LREI to LBVT, LSLM to LBVT and LFSR to LDSY. Trade was not possible for LDSY-LBVT and LOML-LBVT pairs since the spread does not reach the 2 standard deviation threshold. Single trade was possible in the following pairs: LFSR-LBVT and LSLM-LBVT pairs. There was a possibility of two trades in LREI-LBVT pair and LFSR-LDSY pair. The trade was closed when relationship returns to its statistical normal. Opening trade gives investors and traders the opportunity to earn profit from statistical mispricing.

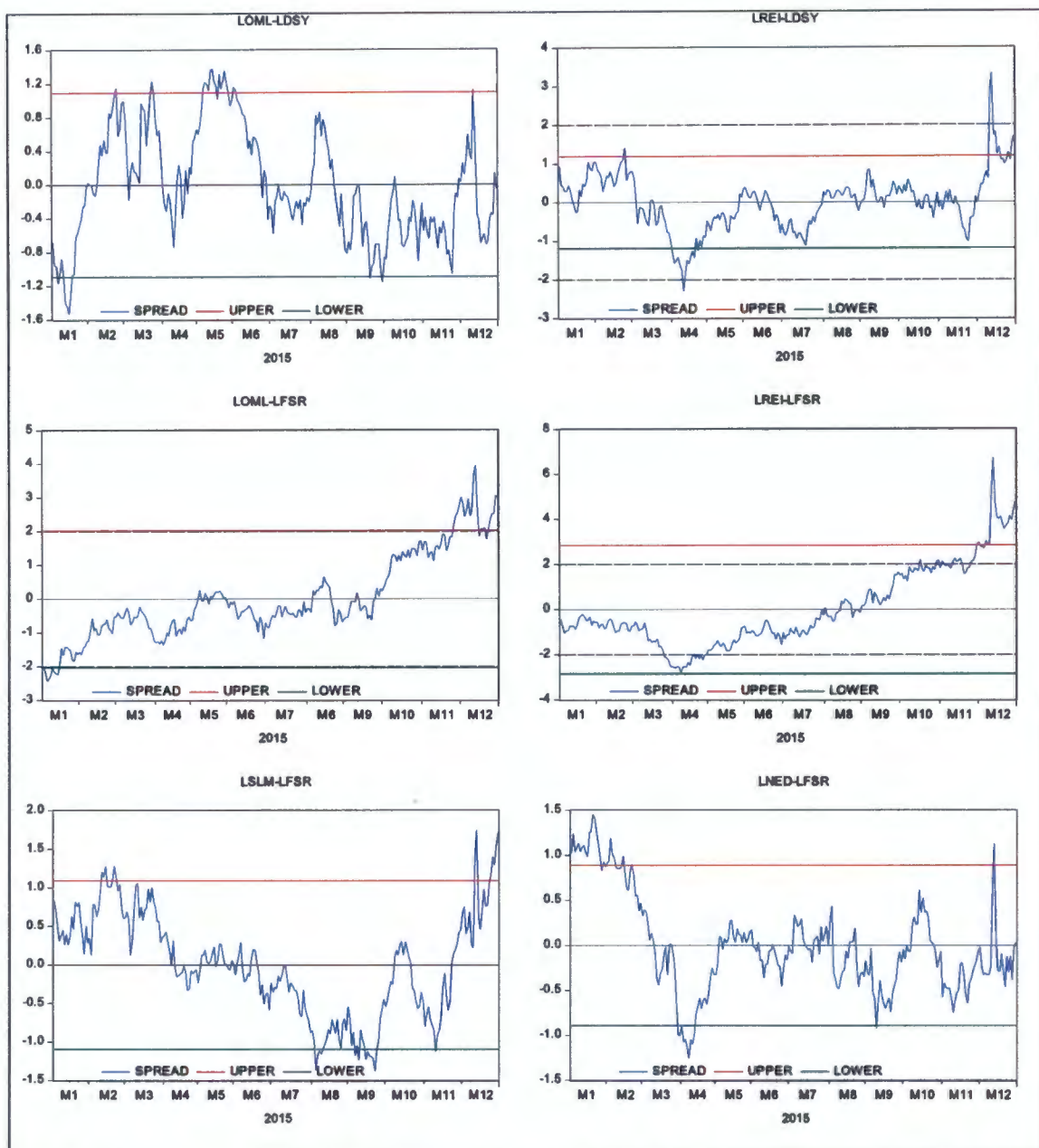


Figure 4.1.13: Spread of LOML-LDSY, LREI-LDSY, LOML-LFSR, LREI-LFSR, LSLM-LFSR and LNED-LFSR pairs with trading thresholds

Figure 4.1.13 depicts that an investor who was determined to trade would simply not trade at all with LOML-LDSY, LSLM-LFSR and LNED-LFSR pairs of stock. An investor would rather hold stocks and wait for future arbitrage opportunities. There were two trades possible for LOML-LFSR, LREI-LFSR and LREI-LDSY pairs. The first trade for LREI-LFSR and LREI-LDSY pairs occurred when the ratio reached -2 standard deviation threshold and the second trade occurred when the ratio reached 2 standard deviation threshold. LFSR and LREI were underperforming relative to LREI and LDSY respectively. Therefore, an investor would

go long on LFSR/LREI while simultaneously going short on LREI/LDSY respectively, and then close trade when relationship returns to its statistical normal.

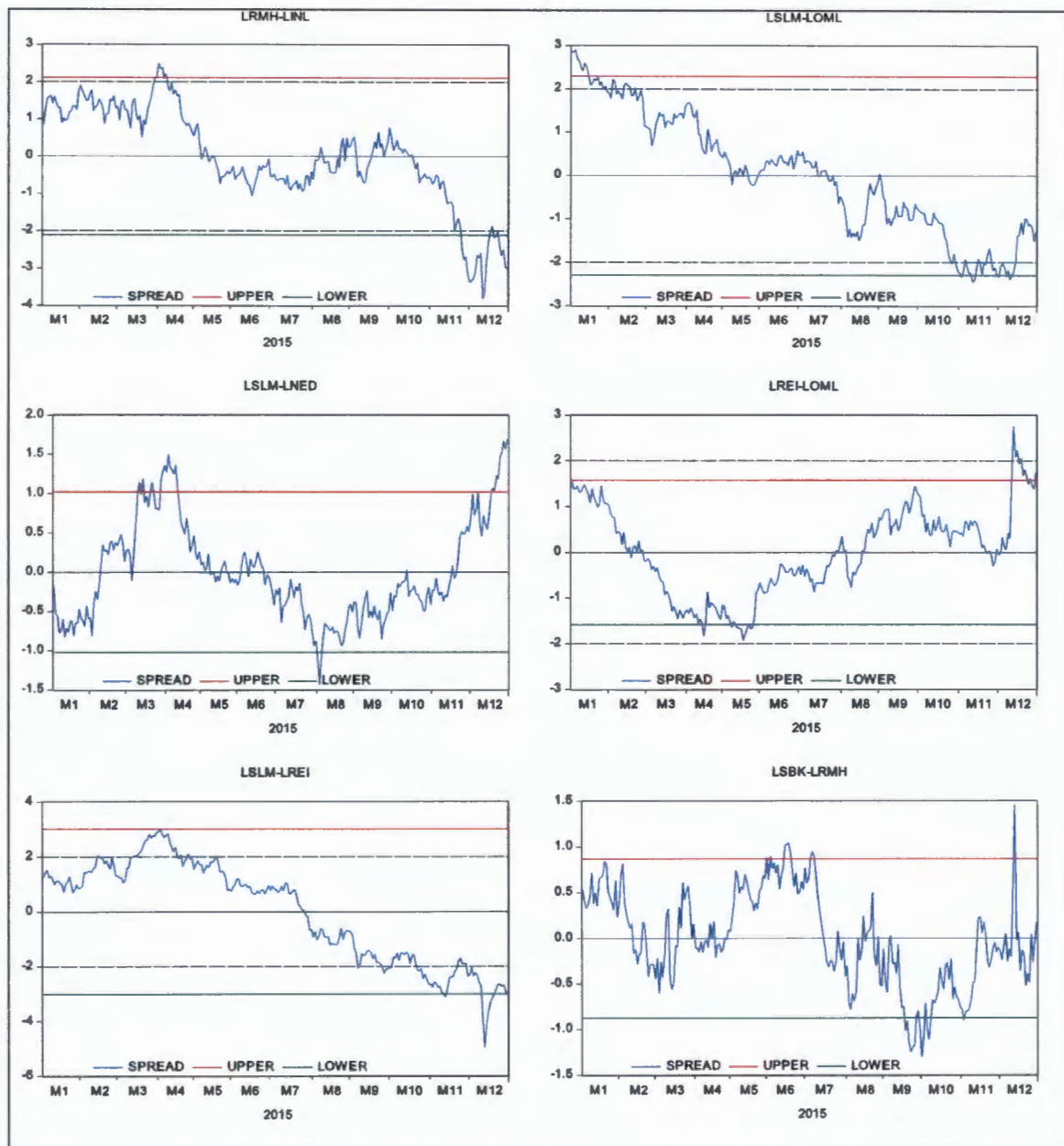


Figure 4.1.14: Spread of LRMH-LINL, LSLM-LOML, LSLM-LNED, LREI-LOML, LSLM-LREI and LSBK-LRMH pairs with trading thresholds

The spread of LRMH- LINL, LSLM-LOML and LSLM-LREI depicts that the first trade occurred when the ratio reaches 2 standard deviations. The second trade also occurred when the ratio of LRMH- LINL, LSLM-LOML and LSLM-LREI reached -2 standard deviation. Single trade occurred when the spread of LREI-LOML reached 2 standard deviation. This

means that LOML was underperforming relative to LREI. Therefore, an investor would have purchased units of LOML and sold units of LREI. The investor would have gone long on LOML while simultaneously going short on LREI, and subsequently close trade when relationship returns to its statistical normal. Trade was not possible in the following pairs: LSLM-LNED and LSBK-LRMH.

4.3 PRELIMINARY DATA ANALYSIS

4.3.1 Analysis on industrial sector

Descriptive statistics give clear characteristics of the data used in this study. The descriptive measures used are mean, median, maximum (max), minimum (min), standard deviation, skewness, kurtosis, Jarque-Bera (JB) test for normality together with their p-values and Ljung-Box (LB) statistics (with their p-values). The results are presented in Table 4.2.1.

Table 4.2.1: Descriptive statistics for log transformed industrial sector

<i>Index</i>	<i>LARI</i>	<i>LMND</i>	<i>LMTN</i>	<i>LNMK</i>	<i>LNTC</i>	<i>LSAP</i>	<i>LSOL</i>	<i>LTKG</i>	<i>LVOD</i>
Mean	9.664	9.342	9.694	7.921	7.629	8.106	10.599	8.124	9.224
Median	9.789	9.149	9.680	7.948	7.573	8.106	10.566	8.143	9.305
Max	10.071	10.422	10.166	8.427	8.389	8.797	11.075	9.048	9.648
Min	8.207	8.566	9.192	7.314	7.127	7.659	10.196	7.084	8.573
Std. Dev.	0.363	0.560	0.243	0.289	0.398	0.243	0.214	0.474	0.289
Skewness	-2.079	0.326	0.023	-0.030	0.440	0.429	0.382	-0.027	-0.721
Kurtosis	6.896	1.656	1.803	1.813	1.745	2.530	2.539	2.282	2.388
JB test	2029.02	139.52	89.69	88.337	146.95	59.756	49.835	32.443	153.268
Prob.	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
LB (15)	20366	21920	21629	21609	22149	20723	21483	22090	21519
Prob.	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Obs	1500	1500	1500	1500	1500	1500	1500	1500	1500

The results presented in Table 4.2.1 revealed that the JB test statistics reject the null hypothesis that variables are normally distributed. The log of African Rainbow Minerals Gold (LARI), Nampak (LNMK), Telkom (LTKG) and Vodacom Group (LVOD) are negatively skewed while the log of Mondi Ltd (LMND), MTN Group (LMTN), Netcare Limited (LNTC), Sappi (LSAP) and Sasol (LSOL) are positively skewed. LARI index have higher peaks as compared to other indexes because its kurtosis value of 6.896 is greater than 3. The values of the skewness and the kurtosis indicate that all the variables have asymmetric distributions. The p-value of the LB statistics shows that there is presence of autocorrelation in all the variables up to lag 15.

4.3.2 Graphical presentation of industrial sector

The graphical presentation of the data gives some idea about the underlying data generating process governing each stock price. The nine variables under industrial sector are presented in Figure 4.2.1.

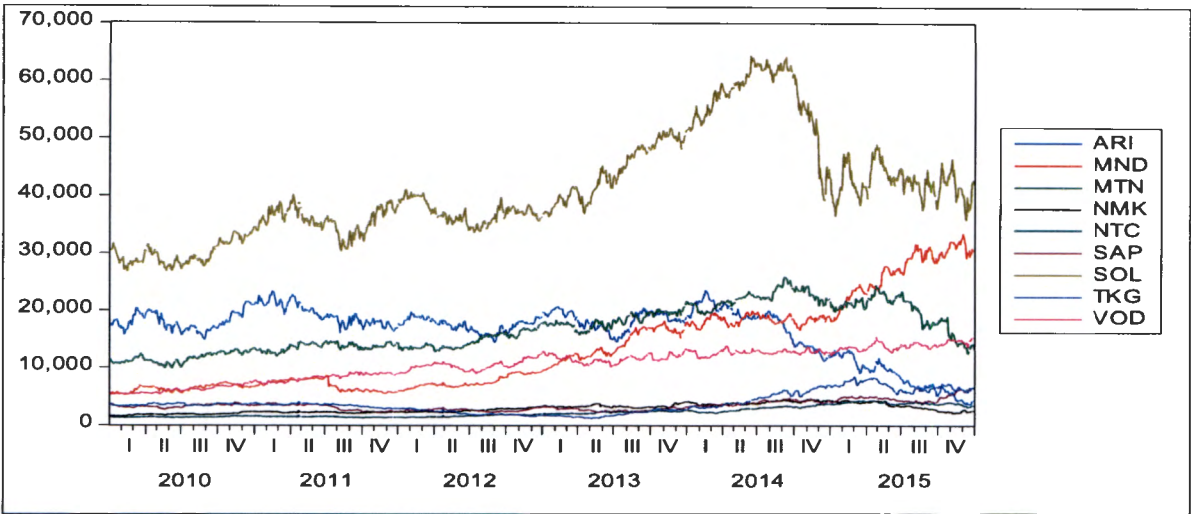


Figure 4.2.1: Time series plot of the historical for volatility of industrial sector at level

Figure 4.2.1 depicts the time series plot of the 9 companies classified as industrial sector in South Africa in this study. The volatility patterns of Sasol (SOL) seem to be increasing progressively over time compared to other indexes. The graphs of African Rainbow Minerals Gold (ARI), Mondi Ltd (MND), MTN Group (MTN), Nampak (NMK), Netcare Limited (NTC), Sappi (SAP), Telkom (TKG) and Vodacom Group (VOD) show a slight trend throughout the sample period. These companies have a slightly similar pattern though MND seems to follow a slight upward trend around the second quarter of 2015. The stock prices do not fluctuate around the mean value and this indicates that the data is nonstationary at level. Therefore, logarithmic transformation was applied to industrial sector variables and the results are presented in Figure 4.2.2.

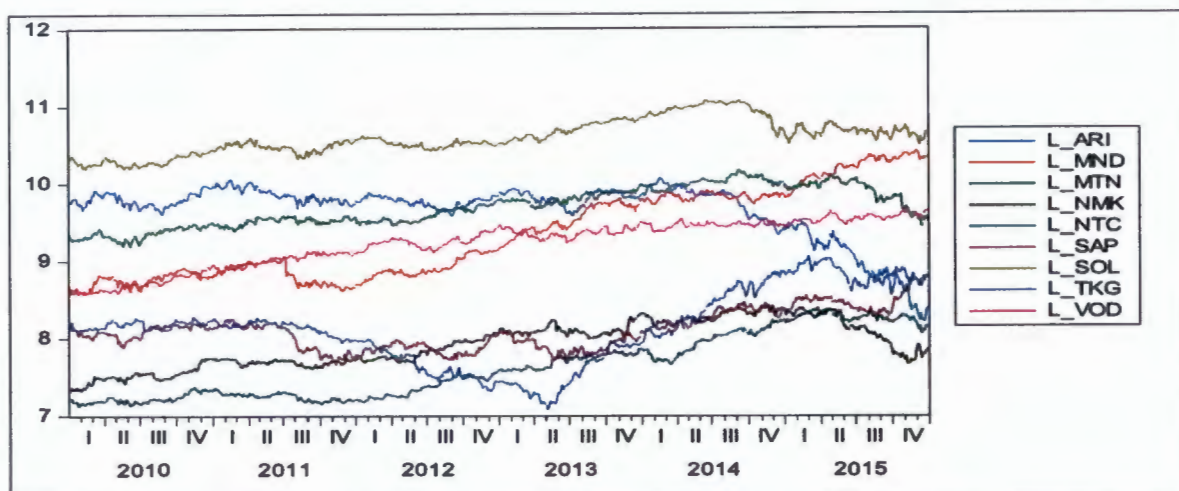


Figure 4.2.2: Time series plot of the historical volatility of industrial sector at logarithm

Figure 4.2.2 shows that the logarithmic industrial sector does not show any sign of mean reversion. Therefore, it is concluded that the logarithm of the industrial sector is not stationary. This means that the data must be differenced. The following, Figure 4.2.3, depicts the log differenced series.

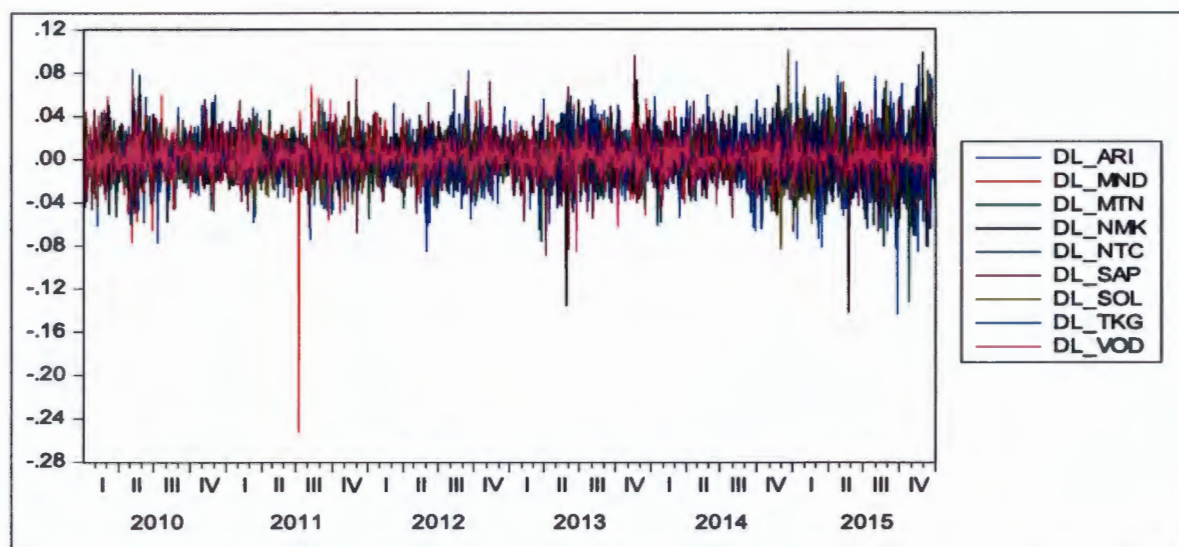


Figure 4.2.3: Time series plot of the historical volatility of industrial sector at logarithm difference

Figure 4.2.3 depicts that the first log differenced indexes fluctuate around the mean value and the variance is relatively stable. Therefore, one can conclude that the data is integrated of order 1, $I(1)$. Table 4.2.2 presents the pairwise correlations for the nine industrial sector companies.

Table 4.2.2: Pairwise correlation of the industrial sector

<i>Correlation</i>	<i>MND</i>	<i>MTN</i>	<i>NMK</i>	<i>NTC</i>	<i>SAP</i>	<i>SOL</i>	<i>TKG</i>	<i>VOD</i>	<i>ARI</i>
MND	1.000								
MTN	0.709	1.000							
NMK	0.619	0.947	1.000						
NTC	0.941	0.796	0.747	1.000					
SAP	0.748	0.402	0.348	0.743	1.000				
SOL	0.580	0.864	0.804	0.577	0.311	1.000			
TKG	0.702	0.440	0.371	0.779	0.867	0.261	1.000		
VOD	0.844	0.831	0.795	0.843	0.480	0.713	0.433	1.000	
ARI	-0.729	-0.261	-0.195	-0.725	-0.634	-0.008	-0.678	-0.532	1.000

Table 4.2.2 presents the pairwise correlation of the companies classified under the industrial sector in this study. Using the cut-off point of 0.90, only two pairs of stocks were selected and were used for further analysis. The selected variable were then tested for stationarity and the results are presented in Table 4.2.3.

4.3.3 Test for stationarity

The two pairs are made of NTC, NMK, MND and MTN. The variables are examined for the presence of unit root. The two tests that are employed are ADF test and the KPSS test. The results are summarised in Table 4.2.3.

Table 4.2.3: Unit root tests results for the industrial sector

<i>Index</i>	<i>ADF test statistic</i>	<i>p-value</i>	<i>KPSS test statistic</i>	<i>p-value</i>
NTC	-0.525	0.884	4.236	0.000
LNTC	-0.352	0.915	4.481	0.000
ΔLNTC	-25.141	0.000**	0.123	0.107**
NMK	-1.762	0.400	3.553	0.000
LNMK	-2.082	0.252	3.699	0.000
ΔLNMK	-39.821	0.000**	0.443	0.440**
MND	0.914	0.996	4.189	0.000
LMND	0.040	0.961	4.466	0.000
ΔLMND	-29.919	0.000**	0.113	0.012
MTN	-1.412	0.578	3.842	0.000
LMTN	-1.570	0.498	3.951	0.000
ΔLMTN	-17.091	0.000**	0.446	0.830**

Note: The ADF and KPSS critical values for the 1%, 5% and 10% significance levels are -3.4345, -2.8633, -2.5677 and 0.7314, 0.4630, 0.3470 respectively. *, ** and *** denotes rejection of the null hypothesis at 1% and 5% and 10% respectively. Lag length was selected by AIC for ADF.

Table 4.2.3 depicts that the variables are nonstationary at level since all the p-values of NTC, NMK, MND and MTN at level are greater than 5% level of significance for ADF and all the p-values of KPSS are less than 5% level of significance. The ADF and KPSS test statistic for LNTC, LNMK, LMND and LMTN are statistically insignificant at 5% level of significance. This means that the null hypothesis could not be rejected. The p-values of ADF test revealed that all the four variables are stationary at first log difference. The p-values for KPSS test also revealed that all the variables except $\Delta LMND$ are stationary at first log difference. This rationalisation is based on the ADF test and it is concluded that the variables are stationary and they are integrated to order 1, $I(1)$. The following subsection tests the relationship between the selected paired indexes.

4.3.4 Relationship between the index pairs

The OLS regression technique is employed to test whether there is a relationship between the two preselected pairs. The regression equation for $\Delta LNMK - \Delta LMTN$ and $\Delta LNTC - \Delta LMND$ is as follows:

$$R_{NMK\ i} = \hat{\beta}_1 + \hat{\beta}_2 R_{MTN\ i} + \hat{\mu}_i \quad (4.6.1)$$

$$R_{LNTC\ i} = \hat{\beta}_1 + \hat{\beta}_2 R_{LMND\ i} + \hat{\mu}_i \quad (4.6.2)$$

The relationships between $\Delta LNMK - \Delta LMTN$ and $\Delta LNTC - \Delta LMND$ are represented in Figure 4.2.4.

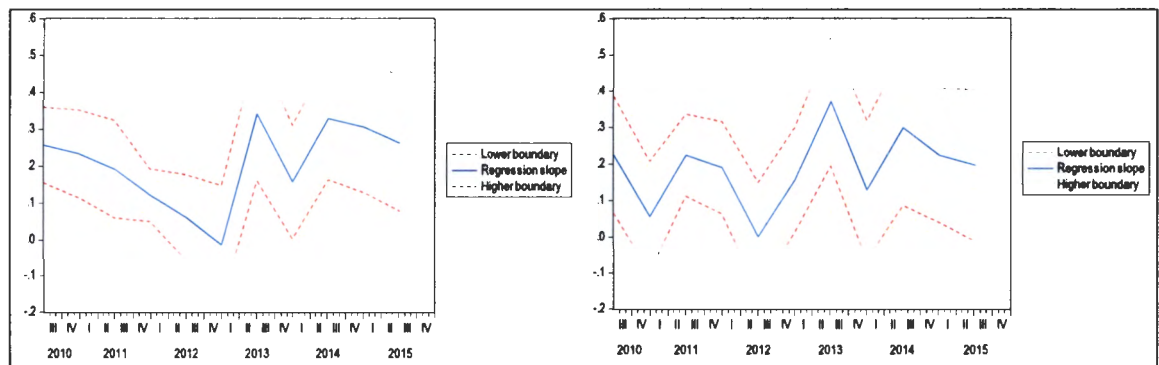


Figure 4.2.4: Beta boundaries for $\Delta LNMK - \Delta LMTN$ and $\Delta LNTC - \Delta LMND$

Figure 4.2.4 presents the obtained $\hat{\beta}_2$ values $\Delta LNMK - \Delta LMTN$ and $\Delta LNTC - \Delta LMND$ from the rolling 125 days regressions starting from 07/05/2010 to 07/06/2015. The graph depicts that the rolling regression slope of $\Delta LNMK - \Delta LMTN$ and $\Delta LNTC - \Delta LMND$ pair falls within the

95% confidence bounds. The rolling coefficients are tested for significance and the results are presented in Table 4.2.4.

Table 4.2.4: Beta boundaries for $\Delta LNMK - \Delta LMTN$ and $\Delta LNTC - \Delta LMND$

<i>Date</i>	$\Delta LNMK - \Delta LMTN$			$\Delta LNTC - \Delta LMND$		
	$\hat{\beta}_2$	t-test	p-value	$\hat{\beta}_2$	t-test	p-value
07/05/2010	0.227	2.83	0.005	0.258	5.039	0.000
12/31/2010	0.057	0.754	0.452	0.234	3.930	0.000
07/05/2011	0.225	4.005	0.000	0.193	2.917	0.004
01/03/2012	0.190	3.004	0.003	0.122	3.430	0.001
07/03/2012	0.002	0.021	0.983	0.062	1.081	0.282
01/02/2013	0.156	2.169	0.032	-0.012	-0.147	0.884
07/03/2013	0.372	4.206	0.000	0.343	3.772	0.000
01/02/2014	0.129	1.349	0.180	0.158	2.052	0.042
07/07/2014	0.300	2.813	0.006	0.330	3.954	0.000
01/05/2015	0.224	2.435	0.016	0.307	3.444	0.001
07/06/2015	0.196	1.89	0.061	0.263	2.861	0.005

Table 4.2.4 depicts the rolling coefficients together with their t-tests and p-values. Most of the regression coefficients were found to be significantly different from zero at a 95% confidence level. Most of the coefficients of $\Delta LNMK - \Delta LMTN$ are significant since their p-values are less than 5% level of significance except for 12/31/2010, 07/03/2012 and 01/02/2014. The results also revealed that most of the rolling coefficients of $\Delta LNTC - \Delta LMND$ pair are significant since the p-values are less than 5% level of significance. The rolling regression shows that the variables are related but it does not indicate whether or not the variables are related in the long run. Therefore, the next step tests where there is a long run relationship between the preselected pairs.

4.3.5 Verification of cointegration

In order to determine whether there is a long relationship between the preselected pair, the EG cointegration test was applied to the two preselected paired indexes. The results are presented in Table 4.2.5.

Table 4.2.5: Engle and Granger cointegration results for LNMK – LMTN and LNTC – LMND

Pairs	OLS Results		Residuals from OLS	
	β_1 Coefficient	t-statistic	ADF (t-statistic)	p-value
LNMK – LMTN	1.133	120.744	-4.873*	0.000
LNTC – LMND	0.686	141.649	-2.505	0.115

Note: *, **, and *** indicates the MacKinnon critical values at 1%, 5% and 10% levels are -3.435, -2.863, and -2.568 respectively. Lag length was selected by SIC.

Table 4.2.5 presents the results of cointegration between the preselected pairs. The results reveal that LNTC-LMND pair is not cointegrated since its t-statistic value of -2.505 is statistically not significant at 1%, 5% and 10% levels. However, the t-statistic value of -4.873 for LNMK-LMTN pair is significant at 1% level. Therefore, it is concluded that there is a long run relationship between LNMK and LMTN. Now that the long run equilibrium has been established, the next step is to estimate the ECM in a bid to measure the short term dynamics between the pairs.

4.3.6 Error Correction Model

Once the cointegration is established, the next step is to reconcile the long run equilibrium with disequilibrium behaviour in the short run using ECM. The ECM results are summarised in Table 4.2.6.

Table 4.2.6: ECM results for LNMK-LMTN

Cointegrated pairs	Stock price	Speed of adjustment	
		γ_1	γ_2
LNMK-LMTN	LNMK	-0.016***	
	LMTN		0.025***

Note: *, **, *** represents the significant level at 10%, 5% and 1%, respectively

The results in Table 4.2.6 reveal that the disequilibrium terms (γ_1 and γ_2) of the cointegrated pair are significantly different from zero, confirming the earlier results of EG cointegration test. It is confirmed that the two variables are cointegrated, and therefore it is plausible to conclude that the short term disequilibrium term would correct the system over the long term equilibrium term. The speed of adjustment is approximately -0.016 and it is significant at 1% level of significance. The value of the ECM coefficient suggests that approximately 1.6 percent of the disequilibrium is corrected for LNMK. The next step is to present the diagnostic results of residual of the selected model.

4.3.7 Model diagnostics

The model is tested for misspecification by checking whether model’s errors are normally distributed or they show any ARCH effect or serial correlation. The results are summarised in Table 4.2.7.

Table 4.2.7: Diagnostic results for LNMK-LMTN

<i>Test</i>	<i>Test statistic</i>	<i>p-value</i>
Breusch-Godfrey (BG)	2.200	0.333
Lagrange Multiplier (LM)	6.346	0.042
Jarque-Bera (JB)	4441.724	0.000
Skewness	-0.635	
Kurtosis	11.337	
ARCH	1.8245	0.177

The results in Table 4.2.7 revealed that the p-value of 0.333 is greater than 5% level of significance therefore it is concluded that there is evidence of serial correlation. The LM p-value of 0.042 is less than 5% level of significance, therefore the null hypothesis is rejected. This means that there is heteroscedasticity. The ARCH LM test was then applied and the results revealed no evidence of autoregressive conditional heteroscedasticity since the p-value of the ARCH LM is greater than 5% level of significance. The value of JB test is 4441.724 with the p-values is 0.000 means that the residuals are not normally distributed. The value of the kurtosis is more than 3, meaning that the residuals are flat tailed. This is in line with the study by Tsay (2005) that reached similar conclusions.

4.3.8 Granger Causality results

Table 4.2.8: Causality between LMTN and LNMK

Null Hypothesis:	Obs	F-Statistic	Prob.
LMTN does not Granger Cause LNMK	1498	2.276	0.034
LNMK does not Granger Cause LMTN		3.427	0.002

The results presented in Table 4.2.8 revealed a bidirectional causality between LMTN and LNMK. The null hypothesis is rejected since the p-value of 0.034 is less than the critical probability p-value of 0.05. This implies that LMTN does Granger cause LNMK and the

alternative hypothesis is also rejected, meaning that LNMK does Granger cause LMTN.

4.3.9 Implementing pairs trading strategy

This section gives a detailed analysis of the empirical results of the implementation of pairs trading strategy. The first step in pairs trading strategy is to find out whether or not the cointegrated paired stocks are historically moving together. The pair is presented in Figure 4.2.5.

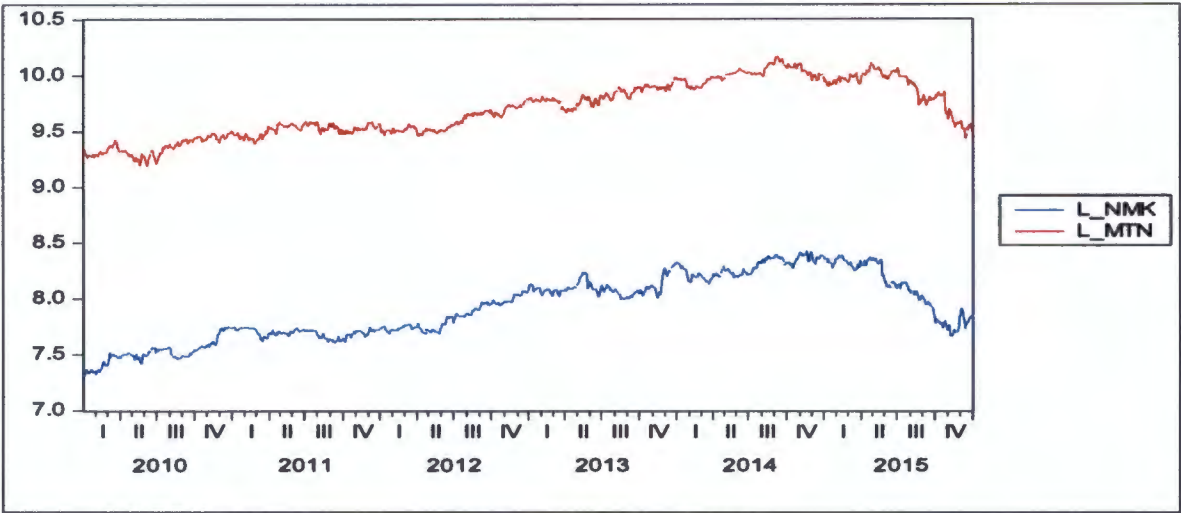


Figure 4.2.5: Historical movement of LNMK – LMTN pair

The LNMK and LMTN price movement shows some convergence throughout the sample period. The next step is to plot the residual series of the LNMK – LMTN pair.

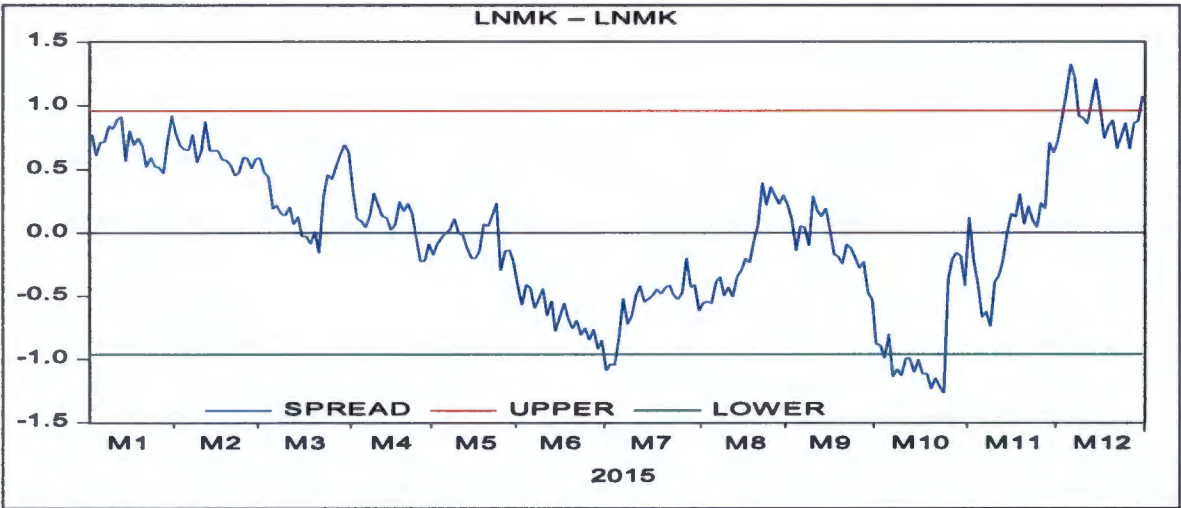


Figure 4.2.6: The residual series of LNMK – LNMK pair

The spread of LNMK and LMTN ranges between -1.5 and 1.5 standard deviation. This means that trade was not possible in 2015 since the spread does not reach the 2 standard deviation threshold.

4.4 PRELIMINARY DATA ANALYSIS

4.4.1 Analysis on mining sector

The descriptive measures summarise the characteristics of the data used. The results of the descriptive statistics are presented in Table 4.3.1.

Table 4.3.1: Descriptive statistics of log transformed mining sector

<i>Index</i>	<i>LACL</i>	<i>LAGL</i>	<i>LAMS</i>	<i>LANG</i>	<i>LBIL</i>	<i>LGFI</i>	<i>LIMP</i>	<i>LKIO</i>	<i>LPPC</i>
Mean	8.320	10.130	10.735	9.948	10.175	8.884	9.415	10.443	7.921
Median	8.299	10.196	10.737	10.160	10.166	9.120	9.512	10.665	7.984
Max	9.366	10.586	11.322	10.571	10.535	9.568	10.078	11.021	8.196
Min	5.670	8.719	9.712	8.906	9.656	8.039	7.751	8.042	7.170
Std. Dev.	0.670	0.286	0.336	0.442	0.151	0.462	0.484	0.579	0.203
Skewness	-1.038	-1.940	-0.368	-0.437	-0.052	-0.248	-1.115	-2.003	-1.257
Kurtosis	4.254	8.251	2.701	1.737	3.230	1.405	3.742	6.809	3.793
JB test	367.722	2664.523	39.481	147.554	3.984	174.497	345.058	1909.462	434.044
Prob.	0.000	0.000	0.000	0.000	0.136	0.000	0.000	0.000	0.000
LB (15)	20570	19218	20694	21660	19440	21880	20929	20508	20343
Prob.	0.000	0.000	0.000	0.000	0.136	0.000	0.000	0.000	0.000
Obs	1500	1500	1500	1500	1500	1500	1500	1500	1500

The results in Table 4.3.1 depict that all the variables exhibit negative skewness. The p-values of JB test statistic indicated that the null hypothesis of normality is rejected for all the variables. This means that all the variables are not normally distributed since the p-values of the JB test are less than 5% level of significance. The data also reveals evidence of leptokurtosis in most of the variables with a kurtosis value greater than 3, which is similar to the findings by Sewell (2011). This means that variables demonstrate a leptokurtosis distribution. This means that the variables are flat tailed. For all the variables, the null hypothesis of no autocorrelation is rejected since the p-values of the LB statistic are less than 5% level of significance.

4.4.2 Graphical presentation of mining sector

The following Figure 4.3.1 presents the historical volatility of the mining sector.

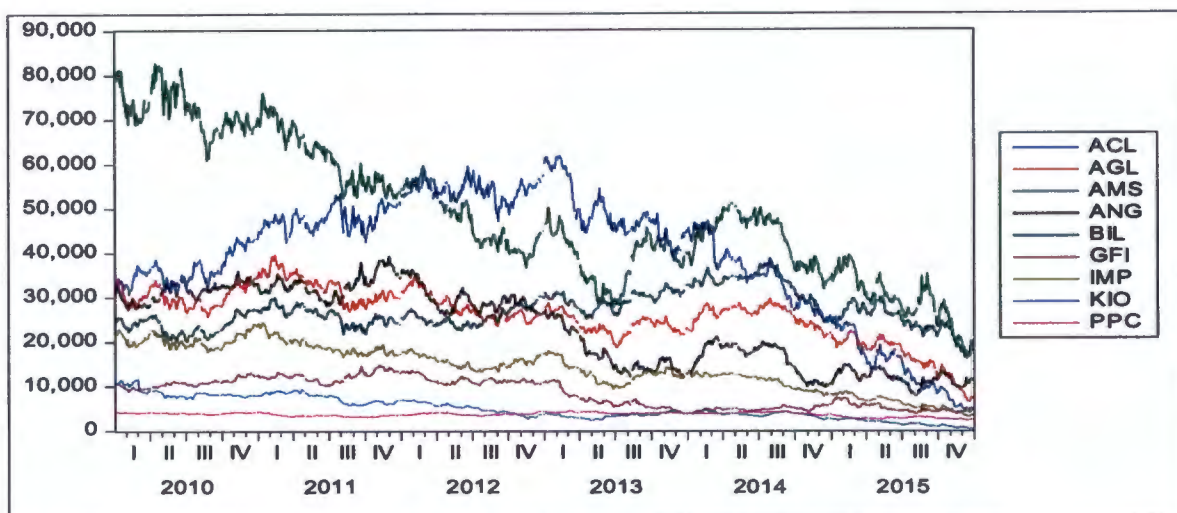


Figure 4.3.1: Time series plot of the historical volatility of mining sector at level

Figure 4.3.1 depicts that the mining sector is not stationary at level. The graph of Anglo American Platinum (AMS) seem to be trending downwards from first quarter of 2010 to last quarter of 2015. The graph of Kumba Iron Ore (KIO) is increasing from first quarter of 2010 until towards the end of 2010, from there it decreases until the end of 2015. Other variables seem to be stable throughout the sample period. It is concluded that all the variables are nonstationary therefore the next step will be to apply log transformation to mining sector. Log transformed results of mining sector is presented in Figure 4.3.2.

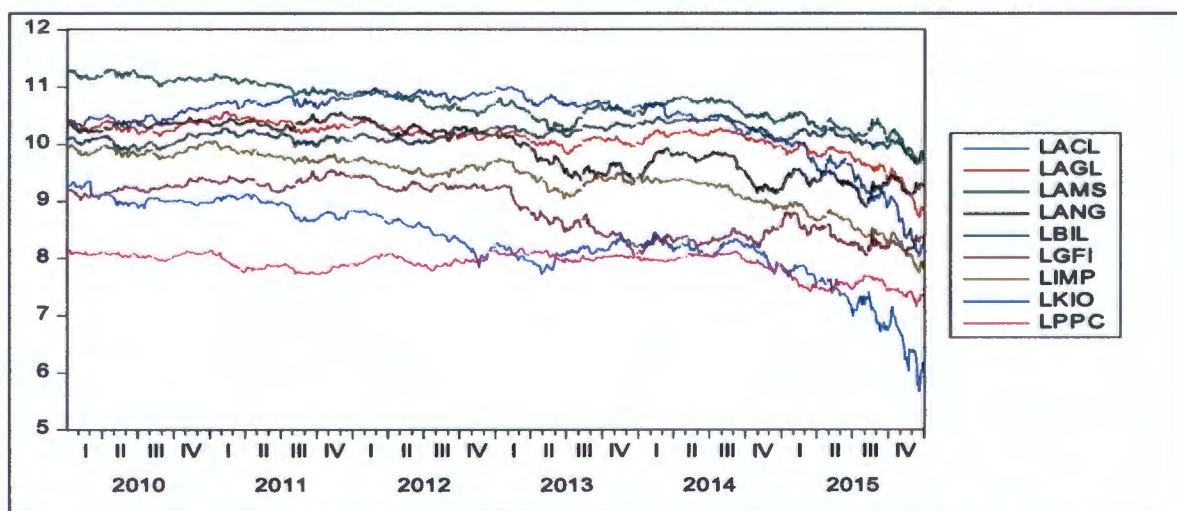


Figure 4.3.2: Time series plot of the historical volatility of mining sector at logarithm

The log transformed data seem to be fluctuating over time. Therefore, differencing was applied to the log transformed data. The following figure presents the log differenced mining sector.

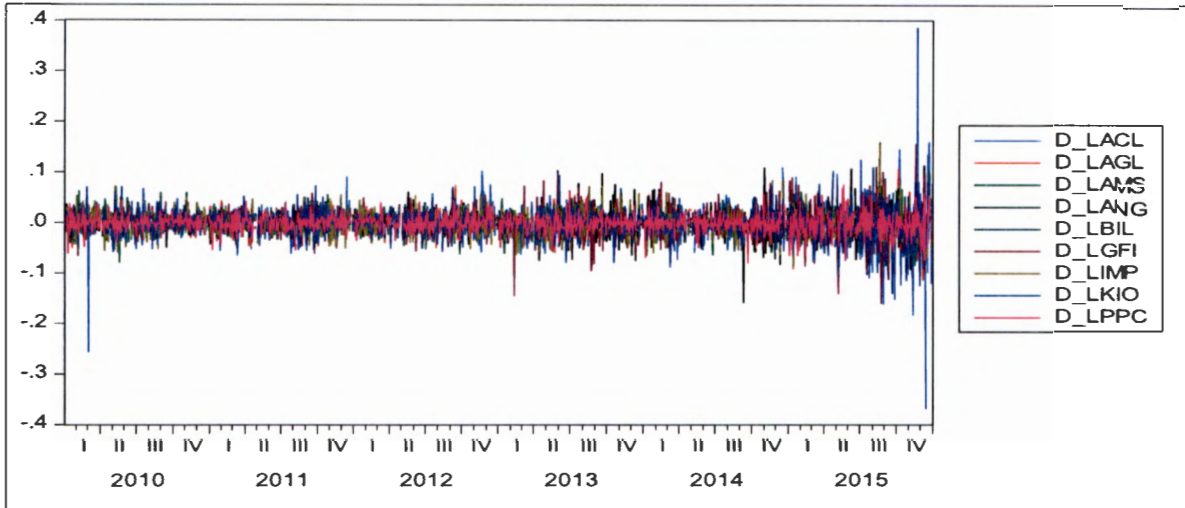


Figure 4.3.3: Time series plot of the historical volatility of mining sector at logarithm difference

Looking at the log differenced mining sector, one could conclude that the variables comprised in the mining sector are stationary. The formal tests for stationarity were applied to confirm stationarity. Table 4.3.2 present the pairwise correlation analysis of the mining sector.

Table 4.3.2: Pairwise correlation of the mining sector

<i>Correlation</i>	<i>ACL</i>	<i>AGL</i>	<i>AMS</i>	<i>ANG</i>	<i>BIL</i>	<i>GFI</i>	<i>IMP</i>	<i>KIO</i>	<i>PPC</i>
ACL	1.000								
AGL	0.842	1.000							
AMS	0.952	0.859	1.000						
ANG	0.819	0.810	0.817	1.000					
BIL	-0.193	0.201	-0.123	-0.262	1.000				
GFI	0.731	0.720	0.699	0.952	-0.381	1.000			
IMP	0.913	0.901	0.925	0.899	-0.078	0.825	1.000		
KIO	0.419	0.681	0.387	0.621	0.253	0.616	0.642	1.000	
PPC	0.482	0.615	0.535	0.389	0.382	0.262	0.618	0.632	1.000

Table 4.3.2 presents the pairwise correlation of the companies classified under mining sector. When using 90% as a cut-off point, only five pairs of stocks were selected and were used for further analysis. The pairs are AMS/ACL, IMP/ACL, IMP/AGL, IMP/AMS and GFI/ANG. Those pairs are then tested for stationarity and the results are presented in Table 4.3.3.

4.4.3 Test for stationarity

The selected paired variables are examined to test the presence of unit root. The results are summarised in Table 4.3.3.

Table 4.3.3: Unit root tests of the mining sector

<i>Index</i>	<i>ADF test statistic</i>	<i>p-value</i>	<i>KPSS test statistic</i>	<i>p-value</i>
AMS	-1.2921	0.6353	4.0495	0.0000
LAMS	-0.3483	0.9152	3.8923	0.0000
ΔLAMS	-29.180	0.0000	0.0840	0.0960
ACL	-1.3368	0.6143	4.2858	0.0000
LACL	2.0189	0.9999	3.8659	0.0000
ΔLACL	-10.1347	0.0000	0.3461	0.0104
IMP	-0.1075	0.9468	4.4098	0.0000
LIMP	2.4307	1.0000	4.0013	0.0000
ΔLIMP	-7.4037	0.0000	0.5137	0.0178
AGL	0.0171	0.9590	3.2887	0.0000
LAGL	2.3792	1.0000	2.9387	0.0000
ΔLAGL	-16.7021	0.0000	0.5298	0.0583
GFI	-0.8266	0.8108	3.8195	0.0000
LGFI	-0.8306	0.8097	3.9003	0.0000
ΔLGFI	-13.2333	0.0000	0.0933	0.4022
ANG	-0.7696	0.8267	4.1620	0.0000
LANG	-0.7428	0.8339	4.1013	0.0000
ΔLANG	-28.3190	0.0000	0.0648	0.2767

*Note: The ADF and KPSS critical values for the 1%, 5% and 10% significance levels are -3.4345, -2.8633, -2.5677 and 0.7314, 0.4630, 0.3470 respectively. *, ** and *** denotes rejection of the null hypothesis at 1% and 5% and 10% respectively. Lag length was selected by AIC for ADF.*

Table 4.3.3 depicts that the AMS, ACL, IMP, AGL, GFI and ANG are nonstationary at level since the p-values for ADF test are greater than 5% level of significance and the p-values for KPSS are less than 5% level of significance. The same conclusion is also reached for all the log transformed variables (LAMS, LACL, LIMP, LAGL, LGFI and LANG). This means that the null hypothesis could not be rejected. The p-values of ADF test revealed that all the variables became stationary after first log difference. The p-values for KPSS test also revealed that all the variables except ΔLACL and ΔLIMP are stationary at log difference. The conclusion is based on the ADF test and it is concluded that all the variables are stationary and they are integrated to order 1, *I*(1). The following subsection examine the relationship between the selected paired indexes.

4.4.4 Relationship between the index pairs

The regression equation for $\Delta LAMS - \Delta LACL$ and $\Delta LIMP - \Delta LACL$ is as follows:

$$R_{AMS\ i} = \hat{\beta}_1 + \hat{\beta}_2 R_{ACL\ i} + \hat{\mu}_i \quad (4.7.1)$$

$$R_{IMP\ i} = \hat{\beta}_1 + \hat{\beta}_2 R_{ACL\ i} + \hat{\mu}_i \quad (4.7.2)$$

The relationship between $\Delta LAMS - \Delta LACL$ and $\Delta LIMP - \Delta LACL$ are represented in Figure 4.3.4.

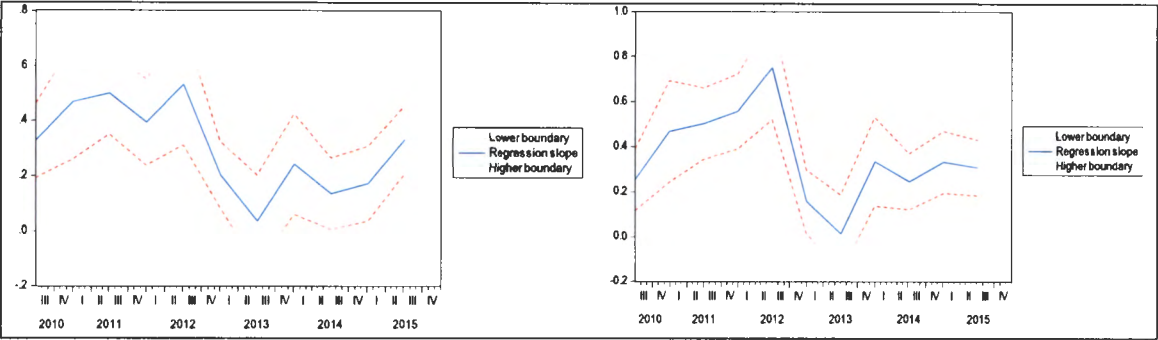


Figure 4.3.4: Beta boundaries for $\Delta LAMS - \Delta LACL$ and $\Delta LIMP - \Delta LACL$

Figure 4.3.4 presents the 95% confidence interval of $\hat{\beta}_2$ values of both $\Delta LAMS - \Delta LACL$ and $\Delta LIMP - \Delta LACL$ obtained from the rolling 125 days regressions starting from 07 May 2010 to 07 June 2015. The rolling regression slope of $\Delta LAMS - \Delta LACL$ and $\Delta LIMP - \Delta LACL$ pair falls between the estimated 95% confidence bounds. The significance test of the rolling coefficients are presented in Table 4.3.3.

Table 4.3.4: Beta boundaries for $\Delta LAMS - \Delta LACL$ and $\Delta LIMP - \Delta LACL$

Date	$\Delta LAMS - \Delta LACL$			$\Delta LIMP - \Delta LACL$		
	$\hat{\beta}_2$	t-test	p-value	$\hat{\beta}_2$	t-test	p-value
07/05/2010	0.328	4.876	0.000	0.255	3.730	0.000
12/31/2010	0.467	4.498	0.000	0.469	4.183	0.000
07/05/2011	0.498	6.780	0.000	0.503	6.383	0.000
01/03/2012	0.394	5.049	0.000	0.557	6.739	0.000
07/03/2012	0.530	4.864	0.000	0.751	6.533	0.000
01/02/2013	0.204	3.416	0.001	0.159	2.267	0.025
07/03/2013	0.037	0.442	0.659	0.016	0.193	0.848
01/02/2014	0.243	2.673	0.009	0.336	3.414	0.001
07/07/2014	0.136	2.104	0.037	0.247	3.995	0.000
01/05/2015	0.173	2.555	0.012	0.333	4.848	0.000
07/06/2015	0.332	5.370	0.000	0.308	5.003	0.000

Table 4.3.4 depicts the rolling coefficients together with their corresponding t-tests and p-values. Most of the coefficients of $\Delta LAMS - \Delta LACL$ are significant since their p-values are less than 5% level of significance except for 07/03/2013. All the coefficients of $\Delta LIMP - \Delta LACL$ are significant since their p-values are less than 5% level of significance except for 07/03/2013.

The regression equation for $\Delta LIMP - \Delta LAGL$ and $\Delta LIMP - \Delta LAMS$ is as follows:

$$R_{IMP\,i} = \hat{\beta}_1 + \hat{\beta}_2 R_{AGL\,i} + \hat{\mu}_i \tag{4.8.1}$$

$$R_{IMP\,i} = \hat{\beta}_1 + \hat{\beta}_2 R_{AMS\,i} + \hat{\mu}_i \tag{4.8.2}$$

The relationship between $\Delta LIMP - \Delta LAGL$ and $\Delta LIMP - \Delta LAMS$ is represented in Figure 4.3.5.

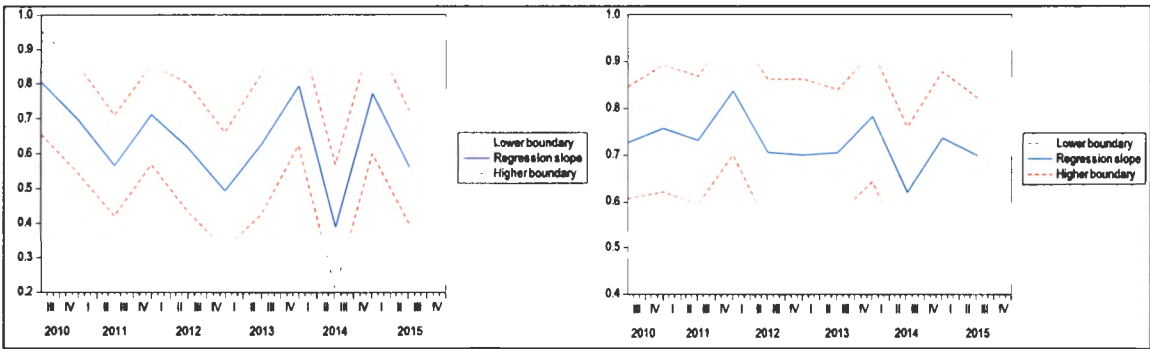


Figure 4.3.5: Beta boundaries for $\Delta LIMP - \Delta LAGL$ and $\Delta LIMP - \Delta LAMS$

Figure 4.3.5 revealed that the rolling regression slope of $\Delta LIMP - \Delta LAGL$ and $\Delta LIMP - \Delta LAMS$ pairs falls within the 95% confidence interval. The test for significance is presented in Table 4.3.5.

Table 4.3.5: Beta boundaries for $\Delta LIMP - \Delta LAGL$ and $\Delta LIMP - \Delta LAMS$

Date	$\Delta LIMP - \Delta LAGL$			$\Delta LIMP - \Delta LAMS$		
	$\hat{\beta}_2$	t-test	p-value	$\hat{\beta}_2$	t-test	p-value
07/05/2010	0.804	10.784	0.000	0.728	12.215	0.000
12/31/2010	0.699	9.073	0.000	0.757	11.208	0.000
07/05/2011	0.566	7.837	0.000	0.732	10.765	0.000
01/03/2012	0.712	10.011	0.000	0.837	12.355	0.000
07/03/2012	0.617	6.738	0.000	0.707	9.083	0.000
01/02/2013	0.494	5.888	0.000	0.701	8.706	0.000
07/03/2013	0.628	6.252	0.000	0.706	10.532	0.000
01/02/2014	0.794	9.394	0.000	0.783	11.343	0.000
07/07/2014	0.389	4.320	0.000	0.621	8.907	0.000
01/05/2015	0.773	8.975	0.000	0.737	10.432	0.000
07/06/2015	0.562	6.965	0.000	0.699	11.319	0.000

Table 4.3.5 depicts the rolling coefficients together with their t-tests and p-values. Every regression coefficient $\Delta LIMP - \Delta LAGL$ and $\Delta LIMP - \Delta LAMS$ was found to be significantly different from zero at a 95% confidence level. It is evident from the table that all the rolling regression coefficients are significant since their associated p-values are less than the 5% significance level.

The regression equation for $\Delta LGFI - \Delta LANG$ is as follows:

$$R_{GFI\ i} = \hat{\beta}_1 + \hat{\beta}_2 R_{ANG\ i} + \hat{\mu}_i \tag{4.9}$$

The relationship between $\Delta LGFI - \Delta LANG$ is represented in Figure 4.3.6.

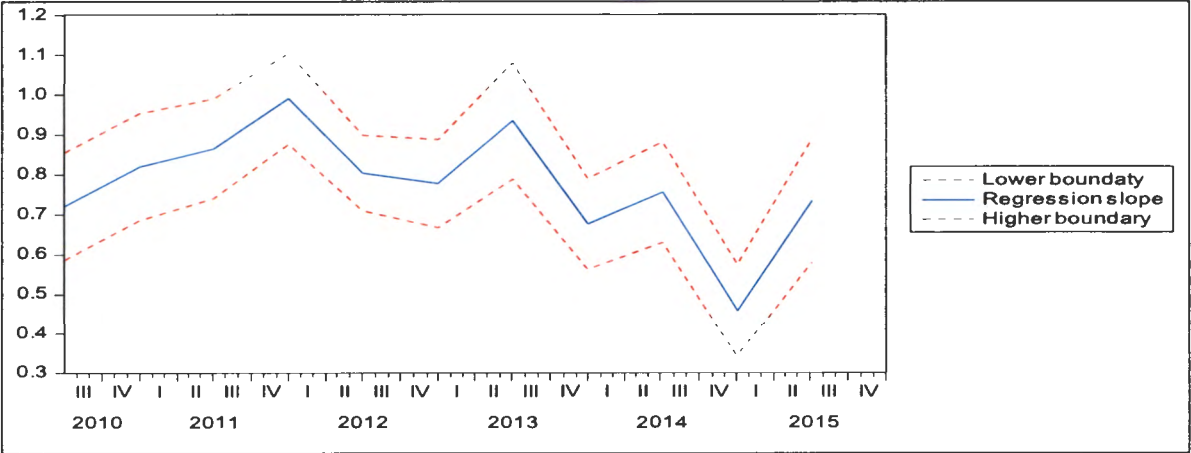


Figure 4.3.6: Beta boundaries for $\Delta LGFI - \Delta LANG$

Figure 4.3.6 presents the 95% confidence limits for $\Delta\text{LGFI} - \Delta\text{LANG}$ pair. The graph depicts that the regression slope of $\Delta\text{LGFI} - \Delta\text{LANG}$ pair falls within the lower limit and the upper limit.

Table 4.3.6: Beta boundaries for $\Delta\text{LGFI} - \Delta\text{LANG}$

<i>Date</i>	$\hat{\beta}_2$	<i>t-test</i>	<i>p-value</i>
07/05/2010	0.722	10.759	0.000
12/31/2010	0.820	12.266	0.000
07/05/2011	0.866	13.888	0.000
01/03/2012	0.991	17.355	0.000
07/03/2012	0.805	17.019	0.000
01/02/2013	0.779	14.126	0.000
07/03/2013	0.935	12.851	0.000
01/02/2014	0.677	11.944	0.000
07/07/2014	0.757	12.052	0.000
01/05/2015	0.460	7.898	0.000
07/06/2015	0.735	9.532	0.000

Table 4.3.6 depicts the rolling regression coefficients together with their t-tests and p-values. Using the 95% confidence limits, all the regression coefficients were found to be significantly different from zero since their p-values are less than 5% level of significance. The rolling regression shows that the variables are related but it does not indicate whether or not the variables are bound to be related in the long run. Therefore, the next step was to test where there is a long run relationship between the preselected pairs using the EG cointegration test.

4.4.5 Verification of Cointegration

The EG cointegration test was applied to the five preselected indexes. The results are presented in Table 4.3.7.

Table 4.3.7: Engle and Granger cointegration results for LAMS – LACL, LIMP – LACL, LIMP – LAGL, LIMP – LAMS and LGFI – LANG

Pairs	OLS Results		Residuals from OLS	
	β_1 Coefficient	t-statistic	ADF (t-statistic)	p-value
LAMS – LACL	0.472	106.857	-4.076*	0.001
LIMP – LACL	0.683	112.930	-3.276*	0.016
LIMP – LAGL	1.570	95.183	-2.266	0.183
LIMP – LAMS	1.322	89.819	-1.572	0.497
LGFI – LANG	0.964	92.309	-1.971	0.300

Note: *, **, and *** indicates the MacKinnon critical values at 1%, 5% and 10% levels are -3.435, -2.863, and -2.568 respectively. Lag length was selected by SIC.

The results presented in Table 4.3.7 show only two pairs are cointegrated since their critical values are statistically significant at 5% level. However, the results reveal that LIMP-LAGL, LIMP-LAMS and LGFI-LANG pairs is statistically insignificant at 5% level of significance. Therefore, the three pairs are not cointegrated. After the long run equilibrium has been established, the next step was to estimate the ECM to measure the short term dynamics of the cointegrated pairs.

4.4.6 Error Correction Model

The cointegration between the variables was established in Section 4.4.5. The next step is to estimate the ECM. The ECM results are summarised in the following Table 4.3.8.

Table 4.3.8: ECM results for LAMS – LACL and LIMP – LACL

Cointegrated pairs	Stock price	Speed of adjustment	
		γ_1	γ_2
LAMS – LACL	LAMS	-0.020**	
	LACL		0.014**
LIMP – LACL	LIMP	-0.008**	
	LACL		0.016**

Note: *, **, *** represents the significant level at 10%, 5% and 1%, respectively

The results in Table 4.3.8 revealed that the disequilibrium terms (γ_1 and γ_2) of the cointegrated pairs are significantly different from zero, which confirms the earlier results of EG cointegration test. The speed of adjustment is approximately -0.020 for the first pair and -0.008 for the second pair. The results also revealed that the speed of adjustment is significant at 5%

level of significance. The next step is to compute the diagnostic tests on the residuals of the model.

4.4.7 Model diagnostics

The model is tested for misspecification by checking whether model’s errors are normally distributed and if they show any ARCH effect or serial correlation. The diagnostics test examine whether the selected model is adequate and efficient. The results are summarised in Table 4.3.9.

Table 4.3.9: Diagnostic results for LAMS – LACL and LIMP – LACL

	<i>LAMS – LACL</i>		<i>LIMP – LACL</i>	
Test	Test statistic	p-value	Test statistic	p-value
Breusch-Godfrey (BG) LM	27.445	0.052	124.808	0.054
Lagrange Multiplier (LM)	28.199	0.000	13.816	0.001
Jarque-Bera (JB)	894.377	0.000	463.938	0.000
Skewness	0.368		0.172	
Kurtosis	6.712		5.704	
ARCH	19.333	0.000	0.129	0.719

The results in Table 4.3.9 reveal that there is no serial correlation in the residual of both the models since their p-values of the BG test are greater than 5% level of significance. The LM p-values of 0.000 and 0.001 are not greater than 5% level of significance, therefore the null hypothesis of heteroscedasticity is rejected. This implies that the error terms of the model have non-constant variance. The p-values of JB test for both pairs are 0.000 for both pairs, implying that the data is not normally distributed. The ARCH LM test for LIMP-LACL pair revealed no evidence of ARCH effect in the model since the p-value of 0.719 is greater than 5% level of significance. The kurtosis value of 6.712 and 5.704 are greater than 3, therefore it is concluded that residuals have a leptokurtic effect.

4.4.8 Granger Causality results

The causality results of the cointegrated pairs are presented in Table 4.3.10.

Table 4.3.10: Causality between LAMS – LACL and LIMP – LACL

Null Hypothesis:	Obs	F-Statistic	Prob.
LACL does not Granger Cause LAMS	1498	7.082	0.001
LAMS does not Granger Cause LACL		2.062	0.128
LACL does not Granger Cause LIMP	1498	1.51767	0.220
LIMP does not Granger Cause LACL		4.41992	0.012

The probability value of 0.001 is less than the critical probability of 5%, therefore the null hypothesis is rejected. This implies that LACL does Granger cause LAMS and the alternative hypothesis is not rejected, meaning that LAMS does not Granger cause LACL. There is a unidirectional long run causality between LACL and LAMS. The results also reveal that LACL does not Granger Cause LIMP since the p-value of 0.220 is greater than 0.05. The alternative hypothesis is rejected and this means that LIMP Granger Cause LACL. Therefore, there is also a unidirectional causality running from LIMP to LACL.

4.4.9 Implementing pairs trading strategy

The first step in implementing pairs trading strategy is to find out whether the paired stocks are historically moving together. The pairs are presented in Figure 4.3.7.

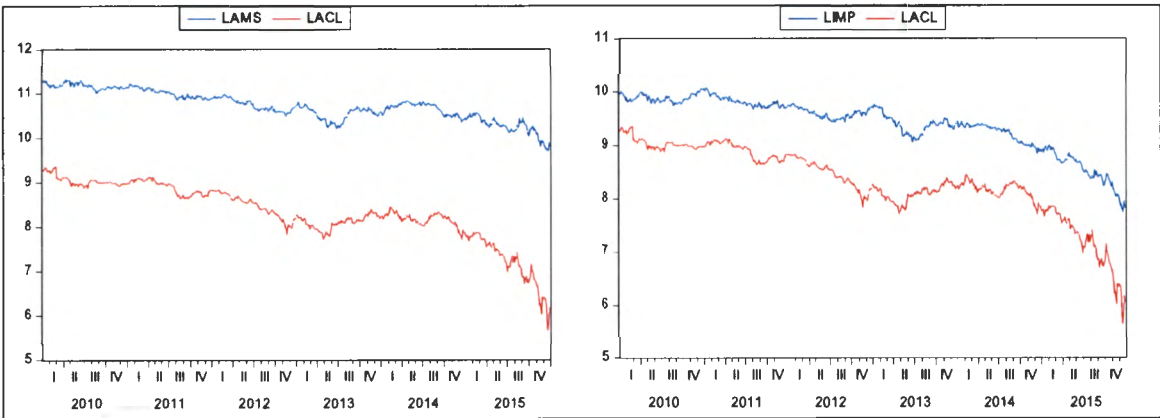


Figure 4.3.7: Historical movement of LAMS – LACL pair and LIMP – LACL pair

Figure 4.3.7 depicts the historic price movement of LAMS – LACL. The price movement seem to have same pattern throughout the sample period. The historic price movement of LIMP and LACL also seem to have a similar pattern throughout the sample period. The price ratios of LAMS – LACL and LIMP and LACL pairs are presented in Figure 4.3.8.

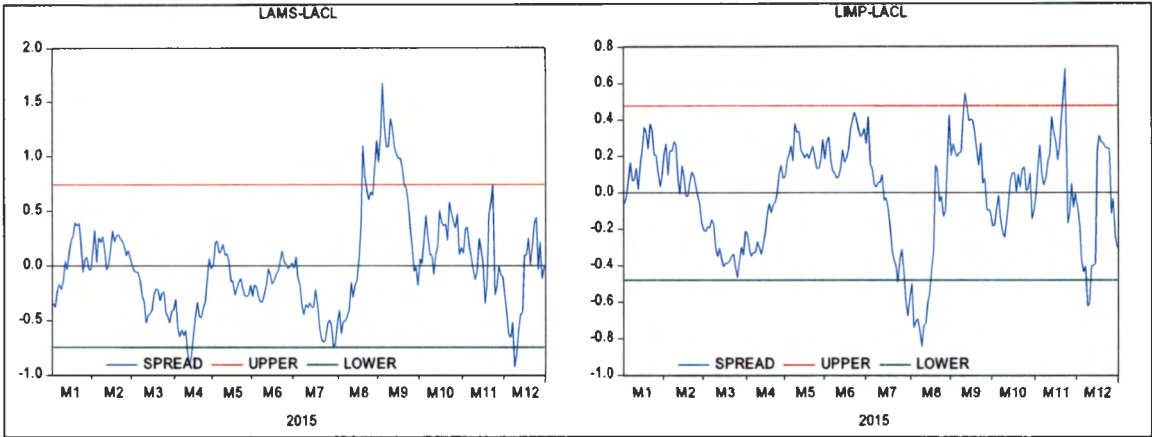


Figure 4.3.8: The residual series of LAMS – LACL and LIMP – LACL pair

The spread of LAMS and LACL ranges between -1.0 and 1.5 standard deviation. This means that there was no arbitrage opportunities between the pairs in 2015 since the spread does not reach the 2 standard deviation threshold. The subsequent figure presents the historic price movement of LIMP – LACL. The spread of LIMP – LACL pair ranges between -1.0 and 0.7 standard deviation. There is also no arbitrage opportunities in LIMP – LACL pair. In conclusion, there were no statistical mispricing for the mining sector. Investors would rather invest elsewhere.

4.5 PRELIMINARY DATA ANALYSIS

4.5.1 Analysis on retail sector

Descriptive measures gives clear understanding of the characteristics of the data used in the study. Financial time series is characterised by stylised facts. Stylised facts of the study are derived from the descriptive measures presented in Table 4.4.1.

Table 4.4.1: Descriptive statistics of log transformed retail sector

<i>Index</i>	<i>LAPN</i>	<i>LBTI</i>	<i>LPIK</i>	<i>LSAB</i>	<i>LSHP</i>	<i>LTBS</i>	<i>LTFG</i>	<i>LTRU</i>	<i>LTSG</i>	<i>LWHL</i>
Mean	9.720	10.691	8.455	10.576	9.509	10.146	9.271	8.954	9.371	8.551
Median	9.697	10.729	8.414	10.597	9.610	10.221	9.301	8.971	9.365	8.767
Max	10.699	11.377	8.836	11.478	9.935	10.595	9.898	9.359	9.926	9.277
Min	8.780	10.009	8.169	9.893	8.773	9.729	8.594	8.328	8.835	7.467
Std. Dev.	0.597	0.373	0.150	0.430	0.278	0.228	0.271	0.189	0.266	0.504
Skewness	0.051	-0.208	0.620	-0.008	-0.798	-0.416	-0.216	-0.889	0.325	-0.481
Kurtosis	1.458	1.768	2.523	1.641	2.430	1.894	2.619	3.949	2.588	1.841
JB test	149.157	105.755	110.241	115.387	179.515	119.816	20.752	253.731	37.028	141.810
Prob.	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
LB (15)	22167	21825	20347	21787	21355	21529	21329	20022	21512	21725
Prob.	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Obs	1500	1500	1500	1500	1500	1500	1500	1500	1500	1500

The highest mean value occurs in British American Tobacco PLC (LBTI) (10.691), while Pick’n Pay Stores (LPIK) has the lowest mean value of 8.455. LBTI, SABMiller (LSAB), Shoprite (LSHP), Tiger Brands (LTBS), The Foschini Group (LTFG), Truworths International (LTRU) and Woolworths Holdings (LWHL) are skewed to the left while Aspen Pharmacare Holdings (LAPN), LPIK and The Spar Group (LTSG) are skewed to the right. The kurtosis value for all the variables except LTRU exceeds the threshold of 3, implying that the variables are platykurtic. Leptokurtosis distribution is only observed in LTRU only. The data also reveals that all the variables reject the hypothesis of normality. Therefore, it is concluded that all the variables are not normally distributed. The p-values of the LB test statistics indicate that all the variables have a presence of autocorrelation.

4.5.2 Graphical presentation of retail sector

The following Figure 4.4.1 presents the graphical presentation of the retail sector.

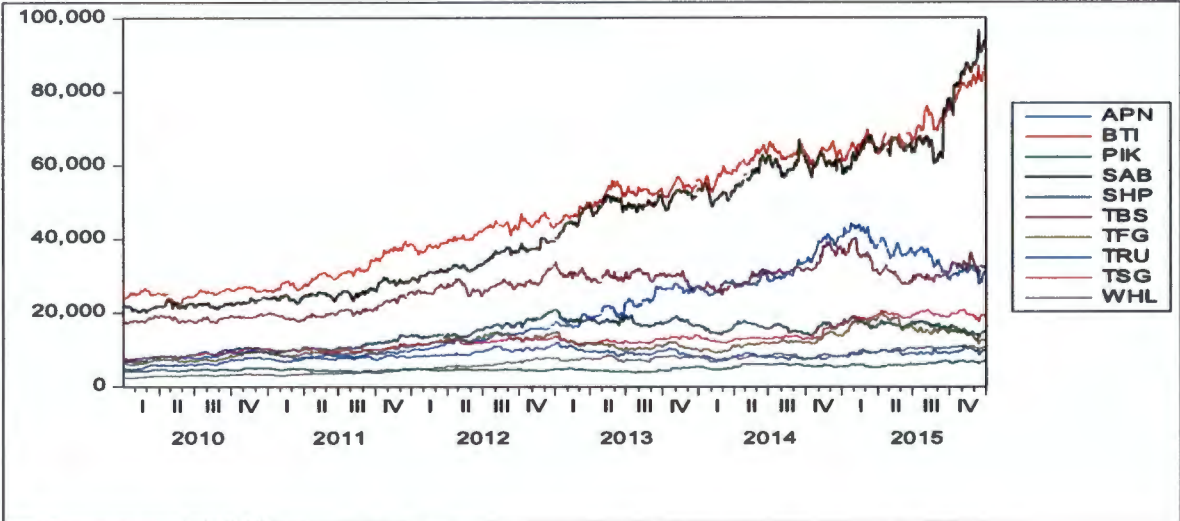


Figure 4.4.1: Time series plot of the historical volatility of retail sector at level

The visual inspection of Figure 4.4.1 depicts that the retail sector is not stationary at level. The graph of BTI and SAB seems to be trending upwards from the start to the end of the sample period. The graph of APN and TBS seems to have a similar pattern from fourth quarter of 2013 until fourth quarter of 2015. Other variables seem to be stable throughout the sample period. It is concluded that all the variables are nonstationary therefore the next step was to apply log transformation to retail sector and the results are presented in Figure 4.3.2.

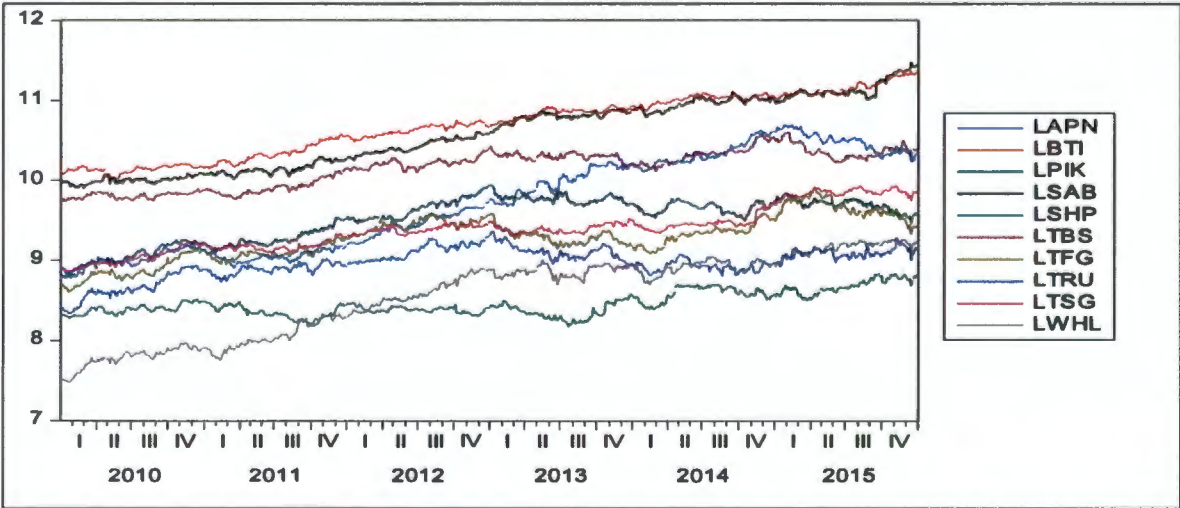


Figure 4.4.2: Time series plot of the historical volatility of retail sector at logarithm

The log transformed data seem to be unstable throughout the sample period. Therefore, differencing was applied to the log transformed data. The log differenced retail sector data is presented in Figure 4.4.3.

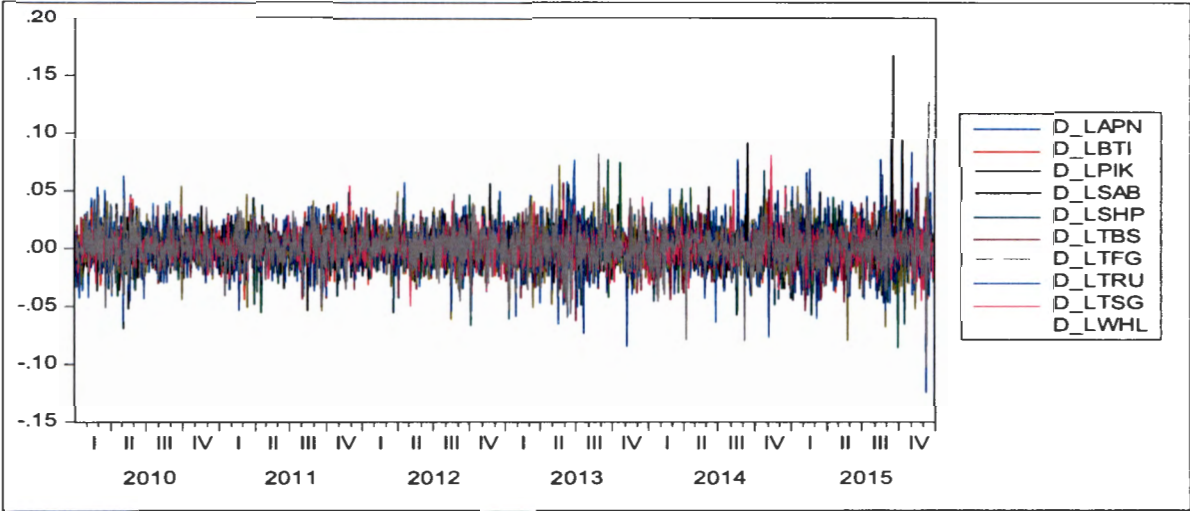


Figure 4.4.3: Time series plot of the historical volatility of retail sector at logarithm difference

The log differenced retail sector presented in Figure 4.4.3 depicts that all the variables under retail sector are stationary. They seem to fluctuate around the mean value. The formal tests for stationarity were then applied to confirm whether the data is stationary or not. The results are presented in Table 4.4.1.

Table 4.4.2: Pairwise correlation of the retail sector

<i>Correlation</i>	<i>APN</i>	<i>BTI</i>	<i>PIK</i>	<i>SAB</i>	<i>SHP</i>	<i>TBS</i>	<i>TFG</i>	<i>TRU</i>	<i>TSG</i>	<i>WHL</i>
APN	1.000									
BTI	0.918	1.000								
PIK	0.707	0.752	1.000							
SAB	0.914	0.988	0.770	1.000						
SHP	0.677	0.737	0.315	0.687	1.000					
TBS	0.846	0.869	0.523	0.842	0.873	1.000				
TFG	0.729	0.744	0.548	0.705	0.745	0.797	1.000			
TRU	0.385	0.540	0.214	0.501	0.833	0.669	0.741	1.000		
TSG	0.854	0.895	0.761	0.884	0.659	0.779	0.887	0.589	1.000	
WHL	0.898	0.969	0.690	0.954	0.834	0.893	0.811	0.663	0.909	1.000

From Table 4.4.2, only six pairs of stocks were selected and used for further analyses. The pairs are BTI/APN, SAB/APN, SAB/BTI, WHL/BTI, WHL/SAB and WHL/TSG, and they were further tested for stationarity and the results are presented in Table 4.4.3.

4.5.3 Test for stationarity

The six selected paired variables are examined to test the presence of unit root. The results are summarised in Table 4.4.3.

Table 4.4.3: Unit root tests of the retail sector

<i>Index</i>	<i>ADF test statistic</i>	<i>p-value</i>	<i>KPSS test statistic</i>	<i>p-value</i>
APN	-0.745	0.833	4.443	0.000
LAPN	-0.906	0.787	4.673	0.000
Δ LAPN	-38.865	0.000	0.162	0.021
BTI	1.129	0.998	4.722	0.000
LBTI	-0.254	0.929	4.727	0.000
Δ LBTI	-28.942	0.000	0.034	0.003
SAB	1.789	1.000	4.625	0.000
LSAB	0.370	0.982	4.761	0.000
Δ LSAB	-30.199	0.000	0.122	0.009
TSG	-0.674	0.851	3.863	0.000
LTSG	-0.995	0.757	4.083	0.000
Δ LTSG	-21.721	0.000	0.052	0.065
WHL	-0.682	0.849	4.593	0.000
LWHL	-1.699	0.432	4.515	0.000
Δ LWHL	-23.876	0.000	0.148	0.010

*Note: The ADF and KPSS critical values for the 1%, 5% and 10% significance levels are -3.4345, -2.8633, -2.5677 and 0.7314, 0.4630, 0.3470 respectively. *, ** and *** denotes rejection of the null hypothesis at 1% and 5% and 10% respectively. Lag length was selected by AIC for ADF.*

Table 4.4.3 depicts that using both ADF test and KPSS test, the original data and the log transformed data is nonstationary. Therefore, differencing was applied to the log differenced data. The p-values of ADF test are all less than 5% level of significance. Therefore, it is concluded that all the variables are stationary after first log difference. The p-values for KPSS test also revealed that only Δ LTSG is stationary at log difference while all the other variables are nonstationary. Since there is a contradiction between two tests of unit root, the conclusion of the study is based on at least one test which is the ADF test and it is concluded that all the variables are stationary. This means that all the variables are integrated to order 1, $I(1)$. The following subsection examines the relationship between the selected paired indexes.

4.5.4 Relationship between the index pairs

The regression equation for $\Delta\text{LBTI} - \Delta\text{LAPN}$ and $\Delta\text{LSAB} - \Delta\text{LAPN}$ are as follows:

$$R_{\text{BTI}i} = \hat{\beta}_1 + \hat{\beta}_2 R_{\text{APN}i} + \hat{\mu}_i \quad (4.10.1)$$

$$R_{\text{SAB}i} = \hat{\beta}_1 + \hat{\beta}_2 R_{\text{APN}i} + \hat{\mu}_i \quad (4.10.2)$$

The relationship between $\Delta\text{LBTI} - \Delta\text{LAPN}$ and $\Delta\text{LSAB} - \Delta\text{LAPN}$ is represented in Figure 4.4.4.

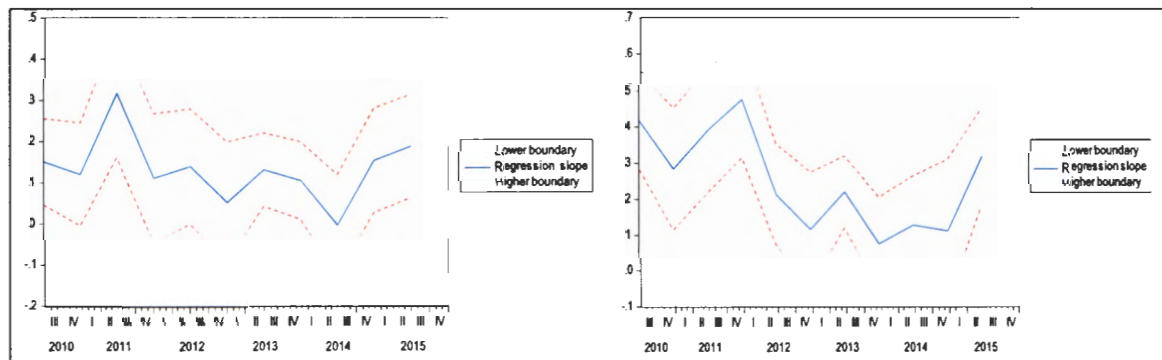


Figure 4.4.4: Beta boundaries for $\Delta\text{LBTI} - \Delta\text{LAPN}$ and $\Delta\text{LSAB} - \Delta\text{LAPN}$

Figure 4.4.4 presents the obtained $\hat{\beta}_2$ values of $\Delta\text{LBTI} - \Delta\text{LAPN}$ and $\Delta\text{LSAB} - \Delta\text{LAPN}$ from the rolling 125 day regressions starting from 07/05/2010 to 07/06/2015. This graph portrays the $\hat{\beta}_2$ falls within the 95% confidence limits. The rolling coefficients were tested for significance and the results are presented in Table 4.4.3.

Table 4.4.4: Beta boundaries for $\Delta\text{LBTI} - \Delta\text{LAPN}$ and $\Delta\text{LSAB} - \Delta\text{LAPN}$

Date	$\Delta\text{LBTI} - \Delta\text{LAPN}$			$\Delta\text{LSAB} - \Delta\text{LAPN}$		
	$\hat{\beta}_2$	t-test	p-value	$\hat{\beta}_2$	t-test	p-value
07/05/2010	0.151	2.896	0.004	0.418	6.229	0.000
12/31/2010	0.121	1.934	0.055	0.284	3.366	0.001
07/05/2011	0.317	4.071	0.000	0.391	4.532	0.000
01/03/2012	0.111	1.423	0.157	0.476	5.923	0.000
07/03/2012	0.139	1.982	0.050	0.214	3.059	0.003
01/02/2013	0.052	0.709	0.479	0.116	1.454	0.148
07/03/2013	0.132	2.952	0.004	0.220	4.419	0.000
01/02/2014	0.105	2.242	0.027	0.076	1.173	0.243
07/07/2014	-0.003	-0.046	0.963	0.128	1.882	0.062
01/05/2015	0.154	2.436	0.016	0.112	1.125	0.263
07/06/2015	0.189	3.012	0.003	0.319	4.736	0.000

Table 4.4.4 depicts the rolling regression coefficients together with their t-tests and p-values. Most of the regression coefficient of $\Delta\text{LBTI} - \Delta\text{LAPN}$ were found to be significant since their p-values are less than 0.05 level of significance. The table also depicts that most of the regression coefficient of $\Delta\text{LSAB} - \Delta\text{LAPN}$ pair are also significant at 5% level of significance. The regression coefficients for 01/02/2013, 01/02/2014 and 01/05/2015 are highly insignificant.

The regression equation for $\Delta\text{LSAB} - \Delta\text{LBTI}$ and $\Delta\text{LWHL} - \Delta\text{LBTI}$ is as follows:

$$R_{SAB\,i} = \hat{\beta}_1 + \hat{\beta}_2 R_{BTI\,i} + \hat{\mu}_i \tag{4.11.1}$$

$$R_{WHL\,i} = \hat{\beta}_1 + \hat{\beta}_2 R_{BTI\,i} + \hat{\mu}_i \tag{4.11.2}$$

The relationship between $\Delta\text{LSAB} - \Delta\text{LBTI}$ and $\Delta\text{LWHL} - \Delta\text{LBTI}$ is represented in Figure 4.4.5.

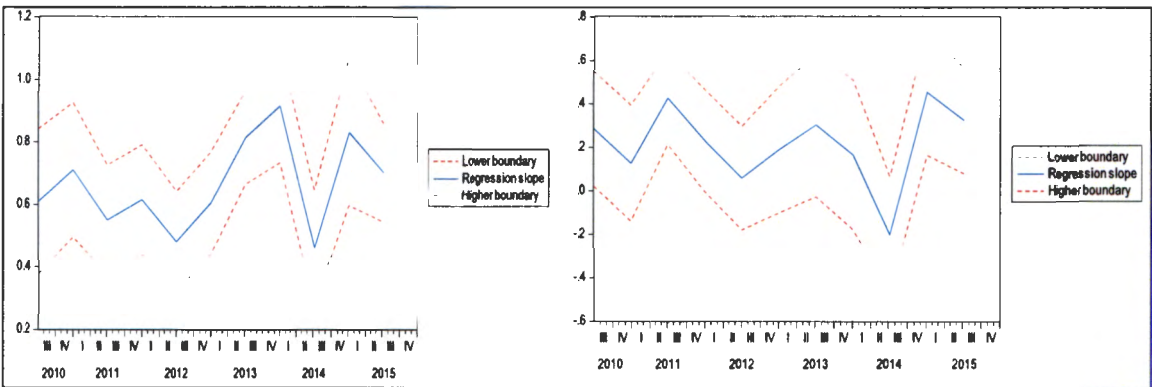


Figure 4.4.5: Beta boundaries for $\Delta\text{LSAB} - \Delta\text{LBTI}$ and $\Delta\text{LWHL} - \Delta\text{LBTI}$

The 95% confidence limits for $\Delta\text{LSAB} - \Delta\text{LBTI}$ and $\Delta\text{LWHL} - \Delta\text{LBTI}$ pair are presented in Figure 4.4.5. The graph shows that the regression slope of $\Delta\text{LSAB} - \Delta\text{LBTI}$ and $\Delta\text{LWHL} - \Delta\text{LBTI}$ pair falls within the 95% confidence limits.

Table 4.4.5: Beta boundaries for $\Delta LSAB - \Delta LBTI$ and $\Delta LWHL - \Delta LBTI$

Date	$\Delta LSAB - \Delta LBTI$			$\Delta LWHL - \Delta LBTI$		
	$\hat{\beta}_2$	t-test	p-value	$\hat{\beta}_2$	t-test	p-value
07/05/2010	0.611	5.289	0.000	0.286	2.172	0.032
12/31/2010	0.711	6.640	0.000	0.128	0.961	0.339
07/05/2011	0.552	6.306	0.000	0.427	4.010	0.000
01/03/2012	0.616	7.052	0.000	0.231	1.979	0.050
07/03/2012	0.483	6.020	0.000	0.060	0.505	0.614
01/02/2013	0.606	7.435	0.000	0.190	1.313	0.192
07/03/2013	0.814	10.959	0.000	0.304	1.844	0.068
01/02/2014	0.915	10.163	0.000	0.167	0.971	0.333
07/07/2014	0.463	5.018	0.000	-0.201	-1.499	0.136
01/05/2015	0.830	7.144	0.000	0.453	3.140	0.002
07/06/2015	0.703	8.933	0.000	0.324	2.658	0.009

The results presented in Table 4.4.5 depict that every regression coefficient is significantly different from zero at a 95% confidence level and all the coefficients of $\Delta LSAB - \Delta LBTI$ are significant since their p-values are less than 5% level of significance. The null hypothesis of $\beta = 0$ is rejected. Table 4.4.5 above also depicts that few of the regression coefficients of $\Delta LWHL - \Delta LBTI$ are significant since their p-values are less than 5% level of significance.

The regression equation for $\Delta LWHL - \Delta LSAB$ and $\Delta LWHL - \Delta LTSG$ is as follows:

$$R_{WHLi} = \hat{\beta}_1 + \hat{\beta}_2 R_{SABi} + \hat{\mu}_i \tag{4.12.1}$$

$$R_{WHLi} = \hat{\beta}_1 + \hat{\beta}_2 R_{TSGi} + \hat{\mu}_i \tag{4.12.2}$$

The relationship between $\Delta LWHL - \Delta LSAB$ and $\Delta LWHL - \Delta LTSG$ are represented in Figure 4.4.6.

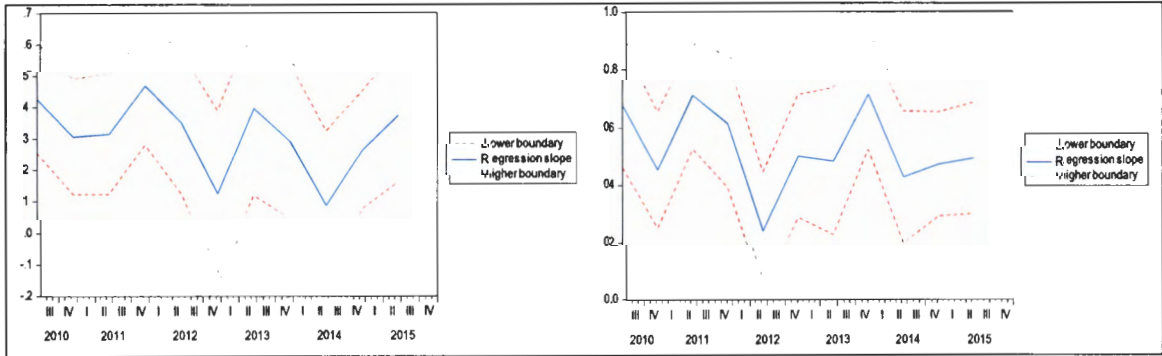


Figure 4.4.6: Beta boundaries for $\Delta LWHL - \Delta LSAB$ and $\Delta LWHL - \Delta LTSG$

The graph in Figure 4.4.6 displays that the regression slope of $\Delta LWHL - \Delta LSAB$ and $\Delta LWHL - \Delta LTSG$ pair falls within the 95% confidence limits.

Table 4.4.6: Beta boundaries for $\Delta LWHL - \Delta LSAB$ and $\Delta LWHL - \Delta LTSG$

Date	$\Delta LWHL - \Delta LSAB$			$\Delta LWHL - \Delta LTSG$		
	$\hat{\beta}_2$	t-test	p-value	$\hat{\beta}_2$	t-test	p-value
07/05/2010	0.427	4.970	0.000	0.680	6.254	0.000
12/31/2010	0.307	3.335	0.001	0.455	4.495	0.000
07/05/2011	0.316	3.267	0.001	0.714	7.673	0.000
01/03/2012	0.469	4.997	0.000	0.615	5.504	0.000
07/03/2012	0.351	3.107	0.002	0.241	2.336	0.021
01/02/2013	0.125	0.936	0.351	0.502	4.692	0.000
07/03/2013	0.397	2.851	0.005	0.485	3.791	0.000
01/02/2014	0.292	2.354	0.020	0.717	7.510	0.000
07/07/2014	0.087	0.728	0.468	0.429	3.740	0.000
01/05/2015	0.264	2.799	0.006	0.474	5.241	0.000
07/06/2015	0.374	3.523	0.001	0.495	5.111	0.000

Table 4.4.6 depicts that most regression coefficients were found to be significantly different from zero at a 95% confidence level and all the coefficients of $\Delta LWHL - \Delta LSAB$ are significant since their p-values are less than 5% level. The regression coefficients of 01/02/2013 and 07/07/2014 are highly insignificant. Every regression coefficient of $\Delta LWHL - \Delta LTSG$ was found to be significantly different from zero at a 95% confidence level and all the coefficients are significant since their p-values are less than 5% level of significance. The rolling regression on Table 4.4.6 shows that the variables are related but it does not indicate whether variables will be related in the long run or not. Therefore, the next step was to test where there is a long run relationship between the preselected pairs using the EG cointegration test.

4.5.5 Verification of Cointegration

The EG cointegration test was applied to the six preselected pairs. The results are presented in Table 4.4.7.

Table 4.4.7: Engle and Granger cointegration results for LBTI – LAPN, LSAB – LAPN, LSAB – LBTI, LWHL – LBTI, LWHL – LSAB and LWHL – LTSG

Pairs	OLS Results		Residuals from OLS	
	β_1 Coefficient	t-statistic	ADF (t-statistic)	p-value
LBTI – LAPN	0.594	119.811	-1.489	0.539
LSAB – LAPN	0.696	145.103	-1.309	0.627
LSAB – LBTI	1.140	244.843	-3.858	0.002**
LWHL – LBTI	1.316	166.503	-3.680	0.005**
LWHL – LSAB	1.125	134.741	-3.257	0.017**
LWHL – LTSG	1.720	84.323	-1.876	0.344

Note: *, **, and *** indicates the MacKinnon critical values at 1%, 5% and 10% levels are -3.435, -2.863, and -2.568 respectively. Lag length was selected by SIC.

The results presented in Table 4.4.7 revealed that only three pairs are cointegrated since their p-values of the ADF test statistic are significant at 5% level of significance. The three pairs are used for further analysis. After the long run equilibrium has been established, the next step was to estimate the ECM to measure the short term dynamics of the cointegrated pairs.

4.5.6 Error Correction Model

The ECM results are summarised in the Table 4.4.8.

Table 4.4.8: ECM results for LSAB – LBTI, LWHL – LBTI and LWHL – LSAB

Cointegrated pairs	Stock price	Speed of adjustment	
		γ_1	γ_2
LSAB – LBTI	LSAB	-0.017**	
	LBTI		0.011**
LWHL – LBTI	LWHL	-0.011**	
	LBTI		0.007**
LWHL – LSAB	LWHL	-0.006*	
	LSAB		0.008**

Note: *, **, *** represents the significant level at 10%, 5% and 1%, respectively

Table 4.4.10 revealed that the disequilibrium terms (γ_1 and γ_2) of the cointegrated pairs are significantly different from zero. The results confirm the earlier the results by EG cointegration test. The disequilibrium terms reveal a bidirectional causality but the results were confirmed in the Granger causality section. The speed of adjustment is approximately -0.017 for the first pair, -0.011 for second pair and -0.006 for the third pair. The results also revealed that the speed

of adjustment are significant at 5% level of significance for first and second pair and 10% for the third pair. The next step is to test the model diagnostics including volatility clustering through ARCH effects in the stock prices.

4.5.7 Model diagnostics

The model is tested for misspecification by checking whether model’s errors are normally distributed or they show any ARCH effect or serial correlation. The results are summarised in Table 4.4.9.

Table 4.4.9: Diagnostic results for LSAB – LBTI, LWHL – LBTI and LWHL – LSAB

<i>Test</i>	<i>LSAB – LBTI</i>		<i>LWHL – LBTI</i>		<i>LWHL – LSAB</i>	
	Test statistic	P-value	Test statistic	P-value	Test statistic	P-value
Breusch-Godfrey (BG) LM	7.545	0.056	3.339	0.188	3.813	0.149
Lagrange Multiplier (LM)	2.230	0.328	11.696	0.003	10.980	0.004
Jarque-Bera (JB)	25071.700	0.000	1171.826	0.000	1479.005	0.000
Skewness	1.902		-0.084		-0.114	
Kurtosis	22.671		7.328		7.861	
ARCH	0.454	0.500	68.140	0.000	78.163	0.000

The results in Table 4.4.9 revealed no a presence of serial correlation for LSAB – LBTI, LWHL – LBTI and LWHL – LSAB pairs. The p-values of 0.003 and 0.004 are less than 5% level of significance, therefore the null hypothesis for homoscedasticity is rejected. This implies that there is a presence of heteroscedasticity. The ARCH test revealed no presence of heteroscedasticity in LSAB – LBTI. There is heteroscedasticity in LWHL – LBTI and LWHL – LSAB pair. The p-values of JB test for the three pairs are 0.000, this implies that the data is not normally distributed. The data also revealed that the residuals have the effect of leptokurtosis, this is given by the kurtosis value greater than 3. This reason behind the insignificance could be due to the high volatility in the datasets.

4.5.8 Granger Causality results

The Granger causality results for the three cointegrated pairs are presented in Table 4.4.10.

Table 4.4.10: Causality between LSAB – LBTI, LWHL – LBTI and LWHL – LSAB

Null Hypothesis:	Obs	F-Statistic	Prob.
LBTI does not Granger Cause LSAB	1498	2.518	0.081
LSAB does not Granger Cause LBTI		1.029	0.358
LBTI does not Granger Cause LWHL	1498	2.592	0.075
LWHL does not Granger Cause LBTI		3.426	0.033
LSAB does not Granger Cause LWHL	1498	0.484	0.616
LWHL does not Granger Cause LSAB		3.867	0.021

Table 4.4.10 presents the causality between the three selected pairs. There is no long run causality between LBTI-LSAB. The results also revealed that LBTI does not Granger Cause LIMP and the results also revealed that LSAB does not Granger cause LWHL. Therefore, there is also a unidirectional causality running from LWHL to LBTI and from LWHL to LSAB. The next step is the implementation of pairs trading strategy.

4.5.9 Implementing pairs trading strategy

The first step in implementing pairs trading strategy is to find out whether or not the paired stocks are historically moving together. The idea of pairs trading is to buy the relatively undervalued stock and sell the relatively overvalued stock, then close trade when the ratio goes back to the mean ration (line 0). The pairs are presented in the following figures.

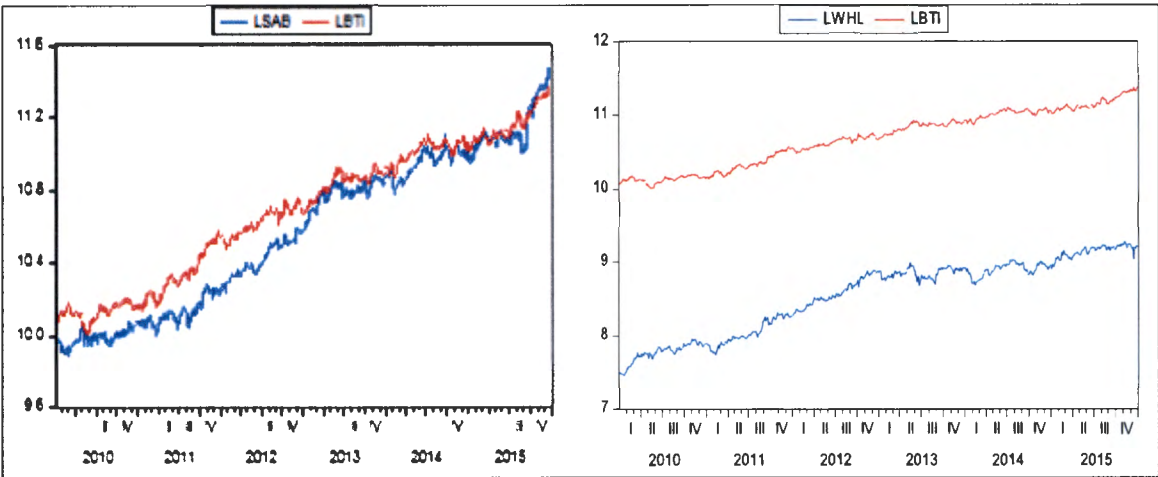


Figure 4.4.7: Historical movement of LSAB – LBTI and LWHL – LBTI pair

Figure 4.4.7 depicts the historic price movement of LSAB – LBTI and LWHL – LBTI pair. The price movement seem to have same pattern throughout the sample period. The figure revealed that there is a close relationship between LSAB – LBTI. The figure also revealed that the historic price movement of LWHL – LBTI seem to have a similar pattern throughout the sample period. The following Figure 4.4.8 presents the price ratios of LSAB – LBTI and LWHL – LBTI pairs.



Figure 4.4.8: Historical movement of LWHL – LSAB pair

Figure 4.4.8 depicts that the historical price movement of LWHL – LSAB seem to have a similar behaviour throughout the sample period. This historical movement in the pairs is important for pair trading.

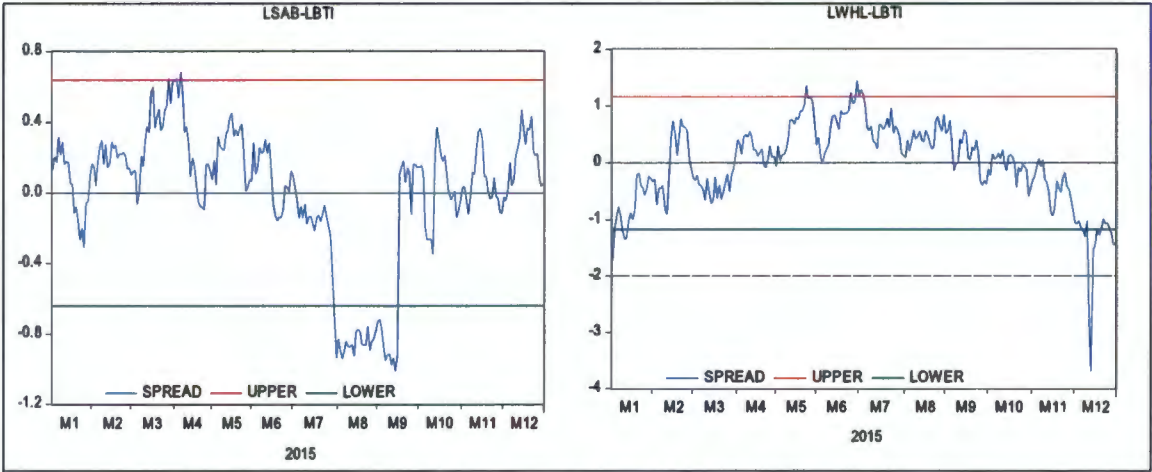


Figure 4.4.9: The residual series of LSAB – LBTI and LWHL – LBTI pair

The spread of LSAB – LBTI revealed that trade was not possible for 2015 since the spread of LWHL – LBTI pair ranges between approximately -4.0 and 1.2 standard deviation. Trade was only possible towards the twelfth month in the LWHL – LBTI pair. This means that LWHL was underperforming relative to LBTI. Therefore, an investor would have bought the units of LWHL and sold the units of LBTI. Implicitly, the investor would have chosen to go long on LWHL while simultaneously going short on LBTI. Trade was closed when relationship returns to its statistical normal.

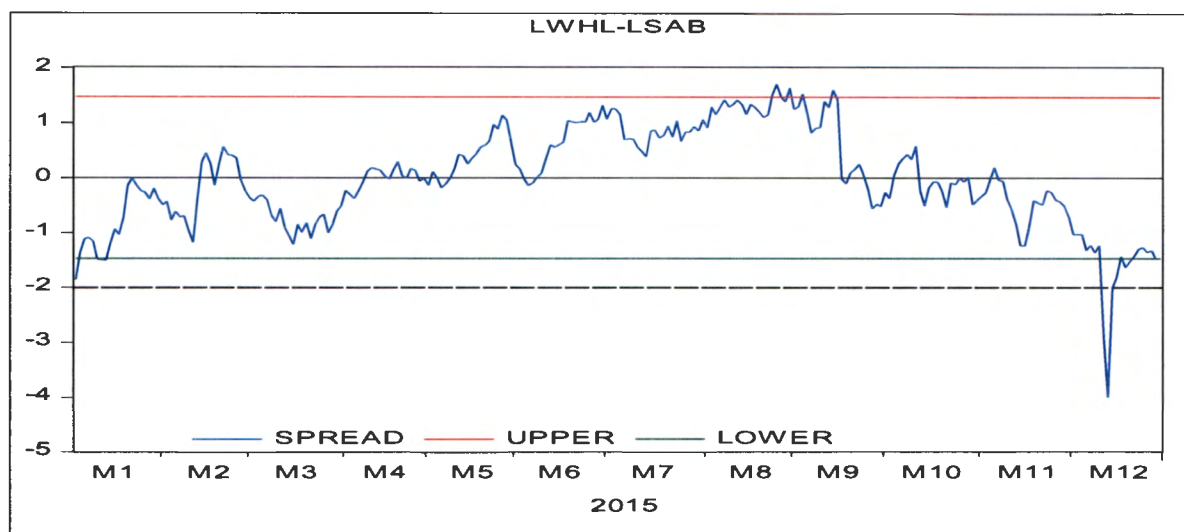


Figure 4.4.10: The residual series of LWHL – LSAB pair

The spread of LWHL – LSAB pair ranges between -4.0 and 1.8 standard deviation. This means that single trade was only possible towards the twelfth month when the ratio reached -2 standard deviation. This means that LWHL was underperforming relative to LSAB. Therefore, an investor had a good opportunity to buy units of LWHL and sell units of LSAB. The investor would go long on LWHL while simultaneously going short on LSAB.

4.6 CHAPTER SUMMARY

This study modelled the volatility of JSE Top 40 index option using statistical arbitrage technique. The study used a daily time series data ranging 04 January 2010 to 31 December 2015 and consist of 1500 observations with one year approximated to have 250 trading days. The technique was employed in four different sectors, namely, financial, industrial, mining and retail. The preliminary analysis was employed to all the four sectors. The variables under each sector were preselected using correlation analysis. The highly correlated pairs were then tested for stationarity using the ADF test and KPSS test. The preselected pairs were tested using the

OLS technique to check whether or not they are related. The preselected pairs were then tested for long run relationship using the Engle and Granger cointegration techniques.

The cointegration technique in this study is also a selection stage prior to the implementation of pairs trading strategy. The ECM was fitted to reconcile the long run equilibrium with disequilibrium behaviour in the short run. The residuals of the models were tested for normality, heteroscedasticity and serial correlation and the associated tests were found to be valid. The Granger technique was also computed to test whether or not adding lags of one variable has a predictive power for another variable. The trading rule used was 2 standard deviations. The pairs trading strategy enables investors to identify pairs of stock that have similar historical movement that could lead to arbitrage opportunities. Several pairs were identified and trade was possible in most of the pairs. The next chapter presents the discussion of results, identifies key conclusions and recommendations.

CHAPTER 5

CONCLUSION AND RECOMMENDATIONS

5.1 INTRODUCTION

This chapter provides summary of the findings of the study, the envisaged contribution of the study, limitations, and draws conclusions in respect of the objectives and suggested recommendations, as well as areas for future research. The remainder of the chapter is structured as follows: Section 5.2 discusses the results of the study in respect of the objectives and gives an indication of how each objective was achieved. Section 5.3 highlights the contribution to the body and horizons of knowledge. Section 5.4 discusses the limitations of the study whereas Section 5.5 draws the conclusions. Section 5.6 discusses the recommendations and areas for future study. Section 5.7 gives a final summary of the thesis.

5.2 DISCUSSION OF RESULTS

This study applied a statistical arbitrage pairs trading strategy to test the efficiency of South African stock markets in pricing relative risk in highly correlated stocks option traded on the JSE. This study further provided a base for future researchers conducting studies on emerging markets, more specifically in South African context. The application of statistical arbitrage trading strategy in the South African context is very limited if at all studies exist. The study tested the possibility of trade in four different sectors trading on the JSE.

The key findings of the study are discussed to address the objectives as follows:

- **To identify pairs of highly correlated index options.**

The strategy of preselecting the pairs is in line with Miao (2014) who used two step approach, namely correlation analysis and cointegration technique, to formulate potential pairs.

Results from the financial sector

Under financial sector, thirty nine pairs of stocks were selected using correlation analysis and they were found to be stationary at first log difference.

Results from the industrial sector

From the industrial sector, out of nine pairs only two stocks were highly correlated and they were found to be integrated to order one, $I(1)$.

Results from the mining sector

With reference to the mining sector, only five pairs of stocks were found to be highly correlated. Those five pairs are AMS – ACL, IMP – ACL, IMP – AGL, IMP – AMS and GFI – ANG and they were also stationary at first log difference just like the financial sector.

Results from the retail sector

Out of ten pairs in the retail sector, only six pairs were selected and they were tested for unit root. Similar to the financial and mining sectors, the unit root tests revealed that the six pairs were stationary at first log difference. A two-step approach was successfully used to select the pairs within each sector. This objective was achieved necessitating the implementation of subsequent steps.

- **To determine whether or not there is a long run relationship (cointegration) between the selected stock indexes.**

Results from the financial sector

Of the thirty nine selected pairs, only eighteen pairs were found to be cointegrated and statistical arbitrage technique was implemented to these cointegrated pairs. The study by Asteriou and Hall (2011) indicated that if variables are cointegrated, their long run relationship is not spurious. This allows for the implementation of the statistical arbitrage pairs trading since this depends on mean reversion (Meki, 2012). The ECM results revealed that the disequilibrium terms (γ_1 and γ_2) of the cointegrated pair are significantly different from zero. The error correction term with a correct sign (-) as presented in Table 4.1.10 were all significant at 5% level.

Furthermore, the diagnostic results revealed that all the cointegrated pairs were not normally distributed since the p-values of the JB test statistic are less than 5% level of significance. This is in line with the views and submissions of Cont (2001) and Yonis (2011).

Results from the industrial sector

The cointegration results revealed that there is a long run relationship between LNMK and LMTN pair. The speed of adjustment was found to be approximately -0.016 and was significant at 1% level of significance. The value of the ECM coefficient suggested that approximately 1.6% of the disequilibrium was corrected for LNMK. Model diagnostics also revealed that the residuals were not normally distributed. The value of the kurtosis was more than 3, meaning that the residuals are flat tailed since the kurtosis value was more than 3. This is supported by the study by Tsay (2005).

Results from the mining sector

The results from the cointegration test revealed that only two pairs, namely LAMS – LACL and LIMP – LACL, were found to be cointegrated. The speed of adjustment was approximately -0.020 for LAMS – LACL pair and -0.008 for LIMP – LACL pair. The results also revealed that the speed of adjustment was significant at 5% level. The p-values of JB test for both pairs were 0.000 for both pairs, implying that the data is not normally distributed. The kurtosis values of 6.712 and 5.704 were greater than 3, therefore it was concluded that residuals have a leptokurtosis effect.

Results from the retail sector

Cointegration test revealed that only three pairs were cointegrated and they were used for further analysis. The speed of adjustment was found to be approximately -0.017 for LSAB – LBTI pair, -0.011 for LWHL – LBTI pair and -0.006 for LWHL – LSAB pair. The speed of adjustment was also found to be significant at 5% level of significance for first and second pair and 10% for the third pair. The JB test revealed that all the three pairs are not normally distributed and the residuals have the effect of leptokurtosis. There was also no serial correlation found. These results are in line with the views by Kat (2003) Ling (2006) and Ang and Chen (2002).

- **To determine the direction(s) of causality between the paired stock indexes.**

Results from the financial sector

Granger causality results revealed that there is a bidirectional causality between LBVT-LDSY and LBVT-LREI and there is a unidirectional causality running from LFSR to LBVT, from LBVT to LOML, from LDSY to LOML, from LDSY to LREI from LREI to LFSR, from

LSLM to LFSR, from LRMH to LINL and also from LNED to LFSR. This means that there is an inter-relationship (one way or two way) between the specified pairs.

Results from the industrial sector

Granger causality results revealed that there is a bidirectional causality between LMTN and LNMK. This means that there is a two way relationship between LMTN and LNMK.

Results from the mining sector

The study also found that there is a unidirectional long run causality running from LACL to LAMS and another unidirectional causality running from LIMP to LACL. The changes in LACL should lead to the changes in LAMS and the changes in LIMP should also lead to the changes in LACL.

Results from the retail sector

The causality test revealed that there is a unidirectional causality running from LWHL to LBTI and from LWHL to LSAB. This means that LWHL has an impact on LBTI in the short run.

- **To identify which pairs of stock are to be traded.**

Pairs trading was then implemented to the selected pairs under financial, industrial, mining and retail sectors. The study used the 2 standard deviation trading rules as previously used by Gatev et al. (2006), Schroder and Smith (2001) and Ruiter (2011).

Results from the financial sector

Single trade was possible in LFSR-LBVT and LSLM-LBVT pairs. There was a possibility of two trades in LREI-LBVT pair and LFSR-LDSY pair. There were two trades possible for LREI-LFSR and LREI-LDSY pairs. The first trade for LREI-LFSR and LREI-LDSY pairs occurred when the ratio reached -2 standard deviation threshold and the second trade occurred when the ratio reached 2 standard deviation threshold. LFSR and LREI were underperforming relative to LREI and LDSY respectively. Therefore, an investor would go long on LFSR/LREI while simultaneously going short on LREI/LDSY and then close trade when relationship returns to its statistical normal. The spread of LRMH- LINL, LSLM-LOML and LSLM-LREI depicts that the first trade occurred when the ratio reaches 2 standard deviation. The second trade also occurred when the ratio of LRMH- LINL, LSLM-LOML and LSLM-LREI reached

-2 standard deviation. Single trade occurred when the spread of LREI-LOML reached 2 standard deviation. Opening trade gives investors and traders the opportunity to earn profit from statistical mispricing.

Results from the industrial sector

Statistical arbitrage was implemented to the LNMK – LMTN pair. The spread of LNMK – LMTN ranged lower than the 2 standard deviation threshold, therefore trade was not possible on the selected pair.

Results from the mining sector

The study further found that there was no statistical mispricing for the mining sector. Therefore there is no possibility of the implementation of pairs trading strategy on the mining sector.

Results from the retail sector

There was only one trade possible for LWHL – LBTI pair. An investor would have chosen to go long on LWHL while simultaneously going short on LBTI. An investor would buy more of LWHL and sell more of LBTI. There was also a single trade possible for LWHL-LSAB pair. This means that an investor had a good opportunity of buying more units of LWHL and selling more units of LSAB.

The results discussed above reveal that the financial sector performed better than the other three sectors. The implementation of pairs trading to historical volatility revealed several possibilities of trade in the financial sector and therefore the investors profit more from trading the identified pairs of stocks in the financial sector. Investors would rather invest elsewhere since there is no possibility of pairs trading in the industrial sector and the mining sector. The results also revealed that there was only one possible trade in the retail sector implying that the possibility of making profit in this sector is very minimal. Note that it is of essence to view all characteristics of the market including but not limited to market volatility as it determines the risk factors in the sector. It is clear from the literature that low volatility yields low returns but it is safer and sustainable.

5.3 CONTRIBUTION TO THE STUDY

According to literature, prevailing methodology on the field suffers deficiency as far as integrated modelling framework which talks to the three modelling steps are concerned. Such steps encompass among others the pre-processing, predictive modelling and decision implementation. This is a critical drawback due to the fact that practically any one of the three steps is extremely contingent upon what literature refers to “statistical mispricings” and “predictive information.” The former is appreciated if and only if they are predictable and the latter becomes useful only if it is effectively exploited. Contrariwise, trading strategy depends on the quality of extrapolative evidence and consequently the predictive model relies on the on-going properties of mispricing dynamics. The current study was motivated by the benefits the methodology of statistical arbitrage strategy offers. The researcher implemented this method successfully and obtained the results that are useful and usable nationally and internationally in related markets.

The study made a contribution to the corpus of knowledge by successfully implementing the statistical arbitrage strategy to categorise statistical mispricing within pairs of stock prices in South Africa. The study also contributes to the scarce empirical evidence of the implementation of pairs trading strategy in South African context. Studies that have been conducted in the South African context do not cover “The implementation of statistical arbitrage technique on the JSE Top 40 Index Option.” Most of the empirical studies have focused on developed countries such as the US and other developing economies where statistical arbitrage strategy has been investigated intensively. This thesis is probably one of the few in the country to apply this internationally approved novel method to the JSE Top 40 Index Option. The study also contributed to the existing literature on historical volatility modelling and stochastic volatility in the field of index option which are traded in emerging markets.

The study constructed a statistical fair price associations as they exist among assets in such a way that the mispricings posed a component that is predictable on the basis of historical stock price movements. The methodology used further allowed the identification of whether or not the deterministic component of the dynamics has some statistical significance as opposed to simple artefact for model building process. The study produced amendable mispricings to statistical arbitrage trading. The results generated from the statistical arbitrage strategy methodology can easily be captured within any fitting predictive model.

The study further contributed to the state of the art of trading by increasing specific aspects of modelling statistical arbitrage methodology and developing tools, such as variable and model selection algorithms, which support the modelling and forecasting of noisy and nonstationary time series. Most importantly, another novelty about this study is the use of tools applied within the context of a modelling framework inspired by techniques from econometric cointegration modelling methods to option pricing. This method form a base upon which the methodology adopted in this study has been developed. The application and implementation of these novel methods was done differently than the traditional one. The application of methods from cointegration analysis is seen as a means to an end. The end in this case is the successfully constructing statistical arbitrage trading strategies. The application of both the econometric and statistical concepts and methods on ordinary financial phenomenon should be noted as a contribution enticing academics in commissioning a study that is multidisciplinary in nature. The data used in the study gave a fresh view on the methods used in the study and also allows South Africa to gauge itself against other emerging markets.

5.4 LIMITATIONS

The data used in this study is secondary and covers closing prices of option and stock for 40 companies traded on JSE and it is obtainable from the JSE. The focus is on the South African context. No surveys, interviews or questionnaires were conducted in this study. The researcher also did not cover other financial variables other than those deemed relevant for the study. The findings of this study may not be generalised as they only apply to the variables used in the study. The findings may not be transferable to other time frames. Due to a dearth of literature on this area, some of the references cited in the study are older than 10 years. This study was an extension of other international ones such as those conducted by Miao (2014), Gatev et al. (2006), Herlemont (2004), Liew and Wu (2013), Schmidt (2008), Schroder and Smith (2001) and Ruiter (2011). It is important to note that searches from the databases such as the National Research Foundation (NRF) and ScienceDirect revealed that there are no studies that used statistical arbitrage approach from the perspective that this research adopted. The current and completed studies do in some cases reference arbitrage, but mostly through the Arbitrage Pricing Theory or through different arbitrage techniques using ARCH and GARCH models.

5.5 CONCLUSION

The study implemented statistical arbitrage pairs trading strategy to four sectors trading on JSE using the Engle and Granger cointegration technique. The main focus of the study was not to determine how much profit can be made from pairs trading strategy but to determine the implementation of pairs trading strategy to JSE top 40 index option. The pairs trading strategy performed better in the financial sector compared to the other three sectors. Specifically, trade was not possible in the industrial and mining sectors. The pairs trading strategy revealed more trading opportunities in the financial sector. Therefore, investors would make profit in trading pairs within the financial sector since there are several possibilities of trade within the sector. The study by Xu (2013) also found that the financial sector also performed better when using the cointegration approach pairs trading strategy. The study took cue from the few of studies as presented in the literature chapter and successfully followed the rule of thumb by focusing on the two standard deviation trading rule when implementing pairs trading strategy. The following authors are amongst those that have used the rule of thumb: Gatev et al. (2006), Schroder and Smith (2001) and Ruiter (2011). The study successfully tested the implementation of the statistical arbitrage pairs trading strategy to JSE top 40 index option of the highly correlated stocks.

5.6 RECOMMENDATIONS FOR FURTHER STUDIES

The study recommends the following areas for future studies:

- The same study could be undertaken to determine how much profit could be made from the implementation of pairs trading strategy and also by taking into account the risks involved in trading in emerging markets such as South Africa. This could help avert making a loss since the main aim of trading is to make profit.
- One can reproduce the same study comparing different methods of cointegration and different trading rules to determine the possibilities of increased trading opportunities. Using different trading rules may enable comparison of different trading possibilities from different thresholds.
- Moreover, a similar study may also be conducted to compare the results of sector restricted pairs and unrestricted pairs using the same data from the JSE. This could help in determining the homogeneity in stock trading pairs within the same sector and equally too from different sectors.

- Another route could be to apply the proposed methodology to high frequency data that would provide more accurate results. Using high frequency data may help increase possible trade in a given day and may lead to having increased number of trades.
- Further study could also make use of the copula approach in implementing pairs trading strategy. The strategy may help in examining the dependency between the stock prices in order to increase the possibilities of trade.
- A comparative study could be undertaken on the implementation and the profitability of pairs trading strategy using cointegration approach, distance approach, stochastic spread approach and the combined forecast approach. The reason for doing a comparative study would determine which method performs better and which one yield more returns.
- The findings of this study can be used in South African stock market together with the fundamental analysis of stocks.

5.7 SUMMARY OF THE THESIS

The thesis was organised into five chapters in order to address the objectives stated in chapter one. Chapter 1 provided the introduction and background of the study, problem statement, rationale of the study, aim and objectives of the study. Chapter 1 also outlined the research questions, significance of the study, scope limitations or delimitations of the study, and definition of terms. Chapter 2 reviewed the relevant literature. The chapter initially discussed the theoretical framework of the study. Different methods of pairs trading were discussed under the theoretical framework. Chapter 2 also examined, discussed and interrogated the empirical literature. Chapter 3 of the study presented the research methodology applied in this study to reach the set objectives while chapter 4 presented the data analysis and interpretation of results. Chapter 5 presented the summary of the findings, drew conclusions in respect of the objectives of the study and suggested recommendations as well as areas for future study.

REFERENCES

- Ackert, L. and Tian, Y. (1999). Efficiency in Index Options Markets and Trading in Stock Baskets. Working paper series, Federal Reserve Bank of Atlanta, 99(5).
- African Securities Exchanges Association (2015). Johannesburg Stock Exchange.
- Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, Vol. 21, pp. 243–247.
- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Caski (Eds.), *Proceedings of the Second International Symposium on Information Theory*, pp. 267–281.
- Akaike, H. (1979). A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. *Biometrika*, Vol. 66, pp. 237–242.
- Akaike, H. (1992). Information theory and an extension of the maximum likelihood principle. In Kotz S. and Johnson, N.L. (Eds.), *Breakthroughs in Statistics*. London, England: Springer-Verlag, Vol. 1, pp. 610–624.
- Alexander, C.O. (1993). The changing relationship between productivity, wages and unemployment in the UK, *Oxford Bulletin of Economics and Statistics*, Vol. 55, pp. 87–102.
- Alexander, C.O. (1999). Optimal hedging using cointegration. *Philosophical Transaction of the Royal Society Series A*, Vol. 357, pp. 2039–2058.
- Alexander, C. (2001). *Market Models: A Guide to Financial Data Analysis*. John Wiley and Sons.
- Alexander, C.O. (2008). *Practical Financial Econometrics*. John Wiley and Sons, pp. 201–250.
- Alexander, C.O and Dimitriu, A. (2015). Indexing and statistical arbitrage: Tracking error or cointegration? *The Journal of Portfolio Management*, pp. 15.
- Alsayed. H. (2014). Essay in Statistical Arbitrage. Financial Economics. University of Southampton Research Repository. ePrint Soton. Unpublished thesis.
- Ammann, M. and Herriger, S. (2002). Relative implied volatility arbitrage with index options. University of St. Gallen, Switzerland. London.

- Andrade, S.C., Di Pietro, V. and Seasholes, M.S. (2005). Understanding the profitability of pairs trading. Unpublished working paper, UC Berkeley, Northwestern University.
- Ang, A. and Chen, J. (2002). Asymmetric correlations of equity portfolios. *Review of Financial Studies*, Vol. 63, pp. 443-494.
- Asteriou, D. and Hall, S. (2011). *Applied Econometrics*. Palgrave Macmillan, New York, 2nd Edition.
- Augustine, C. (2014). Pairs trading: A copula approach. University of Cape Town. Unpublished thesis.
- Avellaneda, M. and Lee, J.H. (2008). Statistical Arbitrage in the US Equities Market. *Quantitative Finance*, Vol. 10, 7, pp.761-782
- Avellaneda, M. and Lee, J.H. (2010). Statistical arbitrage in the US equities market. *Quantitative Finance*, Vol. 10, pp. 1–22.
- Balarezo, J. (2010). International diversification using cointegration and modern portfolio theory. Copenhagen Business School. Unpublished thesis.
- Baronyan, S., Boduroglu, I. and Sener, E. (2010). Investigation of Stochastic Pairs Trading Strategies under different volatility regimes. *Manchester School*, Vol.78, No.s1, pp.114-134.
- Battal, U.K. (2009). Pairs trading in XU100. Istanbul Bilgi University. Unpublished thesis.
- Beckers, S. (1981). Standard deviation implied in option prices as predictors of future stock price variability. *Journal of Banking and Finance*, Vol. 5, 3, pp. 363-381.
- Bianchi, R.J., Drew, M.E. and Zhu, R. (2009). Pairs Trading Profits in Commodity Futures Markets. In *Proceedings of Asian Finance Association 2009 International Conference*, pp. 1–26.
- Bogomolov, T. (2010). Pairs trading in the land down under. Finance and Corporate Governance Conference 2011 Paper. Bundoora, Australia.
- Bolgun, K.E., Kurun, E. and Guven, S. (2009). Dynamic Pairs Trading Strategy for the Companies Listed in the Istanbul Stock Exchange. MPRA Paper No. 19887, pp. 1-23.

- Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics*, Vol. 31, 3, pp. 307-327.
- Bondarenko, O. (2003). Statistical arbitrage and security prices. *Review of Financial Studies*, Vol. 16, pp. 875–919.
- Bowen, D., Hutchinson, M.C. and O’Sullivan, N. (2010). High frequency equity pairs trading: Transaction costs, speed of execution and patterns in returns. *The journal of trading*. Vol. 5, 3, pp. 31-38.
- Brandt, P.T. and Williams, J.T. (2006). Multiple Time Series Models. *Series: Quantitative Applications in the Social Sciences*. Sage Publications Ltd: London.
- Breusch, T.S. and Godfrey, L.G. (1980). A review of recent work on testing for autocorrelation in dynamic economic models. University of Southampton.
- Brockwell, P.J. and Davis, R.A. (2002). *Introduction to Time Series and Forecasting*. 2nd edition, Springer-Verlag, New York, Inc.
- Business Dictionary. (2015). Economic bubble. Retrieved from www.businessdictionary.com/definition/economic-bubble.html.
- Buyuksahin, B., Haigh, M.S. and Robe, M.A. (2008). Commodities and Equities: A Market of One? Working Paper, Washington, DC: US, Commodity Futures Trading Commission.
- Caldas, B., Caldeira, J. and Moura, G. (2014). Is pairs trading performance sensitive to the methodologies? A Comparison. Working paper. Quantitative Finance, ANPEC, pp. 18.
- Caldeira, J. and Moura, G. (2013). Selection of a Portfolio of Pairs Based on Cointegration: A Statistical Arbitrage Strategy. *Brazilian Review of Finance*, Vol. 11, 1, pp. 49–80.
- Chan, E.P. (2009). *Quantitative Trading: How to build your own algorithmic trading business*. New Jersey: Wiley trading series.
- Chan, N.H. (2011). *Time series: Applications to finance with R and S-plus*. John Wiley and Sons.
- Chang, T. (2002). Long-run Benefits from Equity Diversification in two Chinese Share Markets: B-share from Shanghai and Shenzhen Stock exchanges. *The Indian Journal of Economics LXXXII*, pp. 303-310.

Chatfield, C. (2004). *The analysis of time series: An introduction*. 6th edition. New York: Chapman and Hall.

Chiu, M.C. and Wong, H.Y. (2012). Dynamic cointegrated pairs trading: Mean–variance time-consistent strategies. *Journal of Computational and Applied Mathematics*, Vol. 290, pp. 516–534.

CIA World Factbook (2015). South African Economy.

Collis, J. and Hussey, R. (2009). *Business Research: A practical guide for undergraduate and postgraduate students*. 3rd edition, New York: Palgrave Macmillan.

Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance* 1.2, pp. 223-236.

Creswell, J.W. and Tashakkori, A. (2007). Developing publishable mixed methods manuscripts. *Journal of Mixed Methods Research*, Vol. 1, pp. 107-111.

Creswell, J.W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research*. 2nd edition. New York: Pearson Education, Inc. 19.

Creswell, J.W. (2013). *Qualitative inquiry and research design: Choosing among five approaches*. SAGE Publications.

Cummins, R.L. (2010). Pairs trading: Testing the consistency of a statistical arbitrage investment strategy. Working paper, Erasmus University Rotterdam.

De Gooijer, J.G. and Hyndma, R.J. (2006). 25 years of time series forecasting. *International Journal of Forecasting*, Vol. 22, pp. 443–473.

Dickey, D.A. and Fuller, W.A. (1979). Distributions of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of American Statistical Association*, Vol. 74, pp. 427-481.

Dickey, D.A. and Fuller, W.A. (1981). Likelihood Ratio Statistics for Autoregressive Time Series With a Unit Root. *Econometrica*, Vol. 49, pp. 1057-1072.

Do, B., Faff, R. and Hamza, K. (2006). A new approach to modelling and estimation for estimation for pairs trading. Working paper. Monash University.

Dunis, C., Giorgioni, G., Laws, J. and Rudy, J. (2010). Statistical Arbitrage and High Frequency Data with an Application to Eurostoxx 50 Equities. Liverpool Business School, Centre for International Banking, Economics and Finance, pp. 1-31.

Ehrman, D.S. (2006). *The handbook of pairs trading: Strategies using equities, options, and futures*. Wiley. Vol. 240, pp. 167-169

Elliott, G., Rothenberg, T.J. and Stock, J.H. (1996). Efficient tests for an autoregressive unit root. *Econometrica*, Vol. 64, 4, pp. 813-836.

Elliott, R.J., van der Hoek, J. and Malcolm, W.P. (2005). Pairs trading. *Quantitative Finance*, Vol. 5, 3, pp. 271–276.

Enders, W. (1995). *Applied Econometric Time Series*. 2nd edition, John Willey and Sons, pp. 355-400.

Enders, W. (2010). *Applied Econometric Time series*. 3rd edition. New Jersey: John Wiley and Sons.

Engelberg, J., Gao, P. and Jagannathan, R. (2009). An anatomy of pairs trading: the role of idiosyncratic news, common information and liquidity. Unpublished working paper. University of North Carolina, University of Notre Dame, North Western University.

Engle, R.F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica: Journal of the Econometric Society*, pp. 987-1007.

Engle, R.F. and Granger, C.W.J. (1987). Cointegration and error correction: Representation, estimation, and testing. *Econometrica*, Vol. 55, 2, pp. 251–276.

Galenko, A., Popova, E. and Popova, I. (2012). Trading in the presence of cointegration. *The Journal of Alternative Investments*, Vol. 15, 1, pp. 80-97.

Garber, P.M. (1990). Famous first bubbles. *The Journal of Economic Perspectives*, Vol. 2, pp. 35-54.

Garman, E.T. and Fogue, R.E. (2012). *Personal Finance: Real Estate and High-Risk Investment*. South Western: Cengage Learning, pp. 498-499.

Garman, E.T. and Forgue, R.E. (2015). *Personal Finance*. 12th edition. South Western. Cengage Learning.

Gatev, E., Goetzmann, W.N. and Rouwenhorst, K.G. (1999). Pairs trading: Performance of a relative value arbitrage rule. Yale School of Management Working Papers ysm3, pp. 1–34.

Gatev, E., Goetzmann, W.N. and Rouwenhorst, K.G. (2006). Pairs trading: Performance of a relative-value arbitrage rule. *Review of Financial Studies*, Vol. 19, 3, pp. 797–827.

Georgios, K. (2013). Pairs trading in forex markets. A test of the cointegration approach with the EUR/USD and GDP/USD currency pairs. University of Macedonia. Unpublished thesis.

Gephart, R. (1999). Paradigms and research methods. *Research Methods Forum*, Vol. 4, pp. 1-12.

Granger, C.W.J. (1981). Some properties of time series data and their use in econometric model specification. *Journal of Econometrics*, pp. 121-130.

Granger, C.W.J. (1983). Cointegrated variables and error-correcting models, UCSD Discussion Paper, Paper 83-13a.

Granger, C.W.J. (1986). Developments in the Study of Cointegrated Economic Variables. *Oxford Bulletin of Economics and Statistics*, Vol. 48, pp. 213-228.

Granger, C.W.J. and Newbold, P. (1974). Spurious regressions in econometrics. *Journal of Econometrics*, Vol. 2, pp. 111-20.

Gujarati, D.N. (2003). *Basic econometrics*. 4th edition. New York, McGraw-Hill Inc.

Hannan, E.J. and Quinn, B.G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society*, series B, Vol. 41, pp. 190-195.

Hansell, S. (1989). Inside Morgan Stanley's black box. *Institutional Investor* (May), pp. 204-216.

Haque, S.M. and Haque, A.K. (2014). Pairs trading strategy in Dhaka stock exchange: Implementation and profitability analysis. *Asian Economic and Financial Review*, Vol. 4, 8, pp. 1091-1105.

- Hendry, D.F. (1986). Econometric Modelling with Cointegrated Variables. *Oxford Bulletin of Economics and Statistics*, Vol. 48, pp. 201-212.
- Herlemont, D. (2004). Pairs trading, convergence trading, cointegration. YATS Finances and Technologies.
- Hilton, P. and Pedersen, J. (1991). Catalan Numbers, Their Generalization, and Their Uses. *Math. Intel.* Vol. 13, pp. 64-75.
- Hoel, C.H. (2013). Statistical Arbitrage Pairs: Can cointegration capture market neutral profits? (Master's thesis). Bergen: Norwegian School of Economics.
- Hogan, S., Jarrow, R., Teo, M. and Warachka, M. (2004). Testing market efficiency using statistical arbitrage with applications to momentum and value strategies. *Journal of Financial Economics*, Vol. 73, pp. 525-565.
- Hong, G. and Susmel, R. (2003). Pairs trading in the Asian ADR market, Working Paper, University of Houston.
- Huang, R.D., Masulis, R.W. and Stoll, H.R. (1996). Energy Shocks and Financial Markets. *Journal of Futures Markets*, Vol. 16, 1, p.1-27.
- Huck, N. (2009). Pairs selection and outranking: An application to the S&P 100 index. *European Journal of Operational Research*, Vol. 196, 2, pp. 819-825.
- Huck, N. (2010). Pairs trading and outranking: The multi-step-ahead forecasting case. *European Journal of Operational Research*, Vol. 207, pp. 1702–1716.
- Huck, N. and Afabuwo, K. (2015). Pairs trading and selection methods: Is cointegration superior? *Applied Economics*, Vol. 47, 6, pp. 599-613.
- Hull, J.C. (1997). *Options, Futures and Other Derivatives*. 3rd edition, Prentice Hall. New York.
- Hussey, J. and Hussey, R. (1997). *Business research: A practical guide for undergraduate and postgraduate students*. Basingstoke: Macmillan.
- Hwang, A.S. (1996). Positivist and Constructivist Persuasions in Instructional Development. *Instructional Science: An International Journal of the Learning Sciences*, Vol. 24, 5, pp. 343.

Investopedia. (2015). Implied Volatility: Buy Low and Sell High. Retrieved from www.investopedia.com/articles/optioninvestor/08/implied-volatility.asp.

Jacobs, B.I. and Levy, K.N. (1993). Long/short equity investing. *The Journal of Portfolio Management*, Vol. 20, 1, pp. 52-63.

Jacobs, H. and Weber, M. (2011). Losing sight of the trees for the forest? – Pairs trading and attention shifts. Unpublished working paper. University of Mannheim.

Jarque, C.M. and Bera, A.K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters* 6.3, pp. 255-259.

Johansen, S. (1988). Statistical Analysis of Cointegration Vectors. *Journal of Economic Dynamics and Control*, Vol. 12, pp. 231-254.

Johansen, S. and Juselius, K. (1990). Maximum likelihood estimation and inference on cointegration with application to demand for money. *Oxford Bulletin of Economics and Statistics*, Vol. 52, pp. 169-210.

Kanamura, T., Rachev, S. and Fabozzi, F. (2010). A profit model for spread trading with an application to energy futures. *The Journal of Trading*, Vol. 5, 1, pp. 48-62.

Kasa, K. (1992). Common stochastic trends in international stock markets. *Journal of Monetary Economics*, Vol. 29, pp. 95-124.

Kat, H.M. (2003). The dangers of using correlation to measure dependence. *The Journal of Alternative Investments*, Vol. 6, 2, pp. 54-58.

Kiiashko, S. (2012). Index futures efficiency and arbitrage opportunities: Evidence from the Ukrainian Futures Market. Kyiv School of Economics. Unpublished thesis.

Kim, K. (2011). Performance Analysis of Pairs Trading Strategy Utilizing High Frequency Data with an Application to KOSPI 100 Equities, **Working paper, Harvard University.**

Knezevic, A. (2008). Overlapping confidence intervals and statistical significance. Cornell University, Cornell Statistical Consulting Unit. Retrieved from <http://www.cscu.cornell.edu/news/statnews/stnews73.pdf>.

Koronidis, G. (2013). Pairs trading in forex markets: A test of the cointegration approach with the EUR/USD and GBP/USD currency pairs. University of Macedonia, Unpublished thesis.

- Kwiatkowski, D., Phillips, P.C.B., Schmidt, P. and Shin, Y. (1992). Testing the Null Hypothesis of Stationarity against the Alternative of a Unit Root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, Vol. 54, 1-3, pp. 159-178.
- Lee, C.F., Leuthold, R.M. and Cordier, J.E. (1985). The stock market and the commodity futures market: Diversification and arbitrage potential. *Financial Analysts Journal*, Vol. 41, pp. 53-60.
- Liew, R.Q. and Wu, Y. (2013). Pairs Trading: A Copula Approach. *Journal of Derivatives and Hedge Funds*, Vol. 19, 1, pp. 12–30.
- Lin, Y.X., McCrae, M. and Gulati, C. (2006). Loss protection in pairs trading through minimum profit bounds: A cointegration approach, Advances in Decision Sciences. *Journal of Mathematics and Decision Sciences*, Vol 2006, pp. 1-14, DOI 10.1155/JAMDS/2006/73803.
- Ling, H. (2006). Dependence patterns across financial markets: A mixed Copula approach. *Applied Financial Economics*, Vol. 16, 10, pp. 717-729.
- Lo, A.W. (2010). *Hedge funds: An analytic perspective*. Princeton, New Jersey: Princeton University Press, pp. 260-261
- Lucas, A. (1997). Strategic and tactical asset allocation and the effect of long-run equilibrium relations. Series Research Memoranda 0042. VU University Amsterdam, Faculty of Economics, Business Administration and Econometrics.
- Maier, H.R. (2013). What constitutes a good literature review and why does its quality matter? *Environmental modelling and software*, Elsevier Ltd, Vol. 4, pp. 3-4.
- Maree, K. (2012). *First Steps in Research*. Pretoria: Van Schaik Publishers.
- Marwan, M.A. (2015). Examining the relationship between agency costs and stock mispricing: Evidence from the Bahrain stock exchange. *International Journal of Economics, Commerce and Management*. United Kingdom. Vol. 3, 4, pp. 1-35.
- Meki, B. (2012). Examining long-run relationships of the BRICS stock market indices to identify opportunities for implementation of statistical arbitrage strategies. Unpublished thesis.

- Miao, G.J. (2014). High Frequency and Dynamic Pairs Trading Based on Statistical Arbitrage Using a Two-Stage Correlation and Cointegration Approach. *International Journal of Economics and Finance*, Vol. 6, 3, pp. 96-110.
- Mori, M. and Ziobrowski, A.J. (2011). Performance of Pairs Trading Strategy in the U.S. REIT Market. *Real Estate Economics*, Vol. 39, 3, pp. 409–428.
- Muslumov, A.A.Y. (2009). The profitability of pairs trading in an emerging market setting: Evidence from Istanbul stock exchange. *The Empirical Economics Letters*, Vol. 8, 5, pp. 495-500.
- Nath, G.C. and Verma, S. (2003). Study of common stochastic trend and cointegration in the emerging markets: A Case Study of India, Singapore and Taiwan, Research Paper, NSE India.
- Nath, P. (2003). High Frequency Pairs Trading with US Treasury Securities: Risks and Rewards for Hedge Funds, Working Paper, London Business School.
- Ndako, U.B. (2010). Financial development, Economic growth and stock market volatility: Evidence from Nigeria and South Africa. University of Leicester. Unpublished thesis.
- Ng, S. and Perron, P. (1995). Unit root tests in ARMA Models with data dependent methods for the selection of the truncation lag. *Journal of the American Statistical Association*, Vol. 90, pp. 268-81.
- Nicholas, J.G. (2000). *Market neutral investing: long/short hedge fund strategies*. 1st edition. Bloomberg Press, Princeton, 237-240
- Osborne, J.W. (2010). Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research and Evaluation*, Vol. 15, 12, pp. 1-9.
- Oxford Business Group. (2013). The report: South Africa.
- Papadakis, G., and Wysocki, P. (2007). Pairs trading and accounting information. Boston University and MIT Working Paper.
- Pelletier, R. (2007). Improve your trading through correlation analysis. *CSI Technical Journal*, Volume XXV, Vol. 3, pp. 1-7.
- Perlin, M.S. (2007). Evaluation of Pairs Trading Strategy at the Brazilian Financial Market. Working Paper.

- Perlin, M.S. (2009). Evaluation of Pairs Trading Strategy at the Brazilian Financial Market. *Journal of Derivatives and Hedge Funds*, Vol. 15, pp. 122-136.
- Pfaff, B. (2008). Analysis of integrated and cointegrated time series with R. New York: Springer Science Business Media.
- Phillips, P.C. and Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika*, Vol. 75, 2, pp. 335-346.
- Puspaningrum, H., Lin Y.X. and Gulati, C. (2009). Finding the optimal pre-set boundaries for pairs trading strategy based on cointegration technique. Centre for Statistical and Survey Methodology Working Paper Series.
- Rohlf, S. and Zhang, Y. (2016). Culture studies in international business: paradigmatic shifts. *European Business Review*, Vol. 28, 1, pp. 39-62.
- Robson, C. (2002). Real World Research: A Resource for Social Scientists and Practitioner-Researchers, Oxford, Blackwell Publishers, pp. 178.
- Rubin, A. and Babbie, E.R. (2016). Empowerment Series: Research Methods for Social Work. Cengage Learning.
- Rudy, J., Dunis, C. and Laws, J. (2010). Profitable Pair Trading: A Comparison Using the S&P 100 Constituent Stocks and the 100 Most Liquid ETFs. Retrieved from <http://dx.doi.org/10.2139/ssrn.2272791>.
- Ruiter, H.J. (2011). The Performance of a Pairs Trading Strategy in Asian Markets for 2002 to 2009. Unpublished thesis.
- Saunders, M., Lewis, P. and Thornhill, A. (2009). *Research methods for business students*. 5th edition. New York: Pearson Education, Inc. 19.
- Saunders, M., Lewis, P. and Thornhill, A. (2012). *Research methods for business students*. 6th edition. New York: Pearson Education, Inc. 19.
- Saunders, M.N. and Bezzina, F. (2015). Reflections on conceptions of research methodology among management academics. *European Management Journal*, Vol. 33, 5, pp. 297-304.
- Schmidt, A.D. (2008). Pairs trading: A cointegration approach. University of Sydney. Unpublished thesis.

Schroder, S. and Smith, B. (2001). Pairs Trading: A Way to Profit in a Volatile Equity Market. Equity Research Europe.

Schwarz, G. (1978). Estimating the dimension of a Model. *The Annals of Statistics*, Vol. 6, 2, pp. 461–464.

Schwert, W.G. (1990). Stock Market Volatility. *Financial Analysts Journal*, Vol. 46, 3, pp. 23-34.

Sewell, M. (2011). Characterization of financial time series. Research note rn/11/01, UCL, Department of Computer Science, pp 1-35

Song, N. and Chang, S.J. (2003). Nonstationarity and its consequences in modeling the southern timber market. Bugs, budgets, mergers, and fire: Disturbance economics, pp. 241-249.

Ssekuma, R. (2011). *A study of cointegration models with applications*. University of South Africa, South Africa.

Sudak, D. and Suslova, O. (2009). Behavioural Statistical Arbitrage. Unpublished manuscript Thesis University of Lausanne, pp. 1-77.

Syz, J.M. (2008). *Property derivatives: Pricing, Hedging and Application*. John Wiley and Sons Ltd. The Artrium, Southern Gate, Chichester, 12.

Tesfatsion, L. and Kalaba, R. (1989). Time-varying linear regression via flexible least squares. *Computers and Mathematics with Applications*, Vol. 17, 8–9, pp. 1215–1245.

The Economist. (1990). Free Money. 21 July. *Capital Markets*, 16. Vol. 316, Issue 7664.

Tsay, R.S. (2005). *Analysis of financial time series*. John Wiley and Sons: New Jersey.

Ungever, C. (2015). Pairs trading to the commodities futures market using cointegration method. *International Journal of Commerce and Finance*, Vol. 1, 1, pp. 25-38.

Van der Merwe, H. (1996). The Research process: Problem statement and research design. In Garbers, T. G. (Ed) *Effective research in the human sciences. Research management for researchers, supervisors and masters and doctoral candidates*. J. L. van Schaik Publishers. Hatfield, Pretoria.

- Vidyamurthy, G. (2004). *Pairs Trading, Quantitative Methods and Analysis*. Canada: John Wiley and Sons.
- Wakeford, J.J. (2004). The productivity–wage relationship in South Africa: An empirical investigation. *Development Southern Africa*, Vol. 21, 1, pp. 109-132.
- Walls, W.D. (1994). A cointegration rank test of market linkages with an application to the US natural gas industry. *Review of Industrial Organization*, Vol. 9, 2, pp. 181-191.
- Walsham, G. (1995b). The emergence of interpretivism in IS research. *Information Systems Research*, Vol. 6, pp. 376–394.
- Wegner, T. (2016). *Applied business statistics: Methods and excel-based applications*. 4th edition. Juta and Company Ltd, pp. 328-343.
- Wei, W.S. (2006). *Time series analysis: Univariate and multivariate*. Boston: Pearson.
- Weinreich, N.K. (2010). *Hands-on social marketing: A step-by-step guide to designing change for good*. Sage Publications.
- Wilmott, P. (1998). *Derivatives: The Theory and Practice of Financial Engineering*. John Wiley.
- Wong, W.K., Agarwal, A. and Du, J. (2005). Financial integration for Indian stock market: A fractional cointegration approach. National University of Singapore Working Paper No. WP0501.
- Xie, W. and Wu, Y. (2013). Copula-based Pairs Trading Strategy. In *Asian Finance Association (AsFA) 2013 Conference*.
- Xu, X. (2013). Three Essays on Monetary and Financial Economics. University of Alberta. Unpublished thesis.
- Yakop, M. (2011). A comparative analysis of pairs trading. University of Amsterdam. Unpublished thesis.
- Yin, R.K. (2003a). *Case study research, design and methods*. 3rd edition. Thousand Oaks: Sage.
- Yonis, M. (2011). Stock market co-movement and volatility spill-over between USA and South Africa. Unpublished thesis.

Yule, U.G. (1926). Why do we sometimes get nonsense-correlations between time-series? A study in sampling and the nature of time-series. *Journal of the Royal Statistical Society*, Vol. 89, 1, pp. 1-63.

APPENDICES

Appendix A1.1 Rolling regression of the financial sector

Roll:				
Roll type: Fixed window				
Manual specification: D_LDSY D_LBVT				
Estimation command: ROLL(F,WINDOW=125,STEP=125) @ D_LDSY D_LBVT				
Number of subsamples:	11			
Number of coefficients:	1			
Window size:	125			
Step size:	125			
Full sample estimation results:				
Dependent Variable: D_LDSY				
Method: Least Squares				
Date: 09/06/16 Time: 12:14				
Sample (adjusted): 4/02/2010 12/30/2015				
Included observations: 1499 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
D_LBVT	0.440854	0.023720	18.58552	0.0000
R-squared	0.184067	Mean dependent var		0.000949
Adjusted R-squared	0.184067	S.D. dependent var		0.014868
S.E. of regression	0.013430	Akaike info criterion		-5.781955
Sum squared resid	0.270195	Schwarz criterion		-5.778411
Log likelihood	4334.576	Hannan-Quinn criter.		-5.780635
Durbin-Watson stat	2.110850			

Appendix A1.2 Rolling regression of the financial sector

Roll:				
Roll type: Fixed window				
Manual specification: D_LFSR D_LBVT				
Estimation command: ROLL(F,WINDOW=125,STEP=125) @ D_LFSR D_LBVT				
Number of subsamples:	11			
Number of coefficients:	1			
Window size:	125			
Step size:	125			
Full sample estimation results:				
Dependent Variable: D_LFSR				
Method: Least Squares				
Date: 09/06/16 Time: 16:15				
Sample (adjusted): 1/05/2010 12/31/2015				
Included observations: 1499 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
D_LBVT	0.616214	0.025043	24.60603	0.0000
R-squared	0.287057	Mean dependent var		0.000556
Adjusted R-squared	0.287057	S.D. dependent var		0.016793
S.E. of regression	0.014179	Akaike info criterion		-5.673410
Sum squared resid	0.301174	Schwarz criterion		-5.669866
Log likelihood	4253.220	Hannan-Quinn criter.		-5.672089
Durbin-Watson stat	2.165850			

Appendix A1.3 Rolling regression of the financial sector

Roll:				
Roll type: Fixed window				
Manual specification: D_LINL D_LBVT				
Estimation command: ROLL(F,WINDOW=125,STEP=125) @ D_LINL D_LBVT				
Number of subsamples:	11			
Number of coefficients:	1			
Window size:	125			
Step size:	125			
Full sample estimation results:				
Dependent Variable: D_LINL				
Method: Least Squares				
Date: 09/06/16 Time: 16:20				
Sample (adjusted): 1/05/2010 12/31/2015				
Included observations: 1499 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
D_LBVT	0.510657	0.023752	21.49988	0.0000
R-squared	0.235007	Mean dependent var		0.000498
Adjusted R-squared	0.235007	S.D. dependent var		0.015375
S.E. of regression	0.013448	Akaike info criterion		-5.779312
Sum squared resid	0.270910	Schwarz criterion		-5.775768
Log likelihood	4332.595	Hannan-Quinn criter.		-5.777992
Durbin-Watson stat	2.176088			

Appendix A1.4 Rolling regression of the financial sector

Roll:				
Roll type: Fixed window				
Manual specification: D_LNED D_LBVT				
Estimation command: ROLL(F,WINDOW=125,STEP=125) @ D_LNED D_LBVT				
Number of subsamples:	11			
Number of coefficients:	1			
Window size:	125			
Step size:	125			
Full sample estimation results:				
Dependent Variable: D_LNED				
Method: Least Squares				
Date: 09/06/16 Time: 16:25				
Sample (adjusted): 1/05/2010 12/31/2015				
Included observations: 1499 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
D_LBVT	0.527352	0.023247	22.68498	0.0000
R-squared	0.255440	Mean dependent var		0.000281
Adjusted R-squared	0.255440	S.D. dependent var		0.015254
S.E. of regression	0.013162	Akaike info criterion		-5.822285
Sum squared resid	0.259514	Schwarz criterion		-5.818741
Log likelihood	4364.803	Hannan-Quinn criter.		-5.820965
Durbin-Watson stat	2.272311			

Appendix A1.5 Rolling regression of the financial sector

Roll:				
Roll type: Fixed window				
Manual specification: D_LOML D_LBVT				
Estimation command: ROLL(F,WINDOW=125,STEP=125) @ D_LOML D_LBVT				
Number of subsamples:	11			
Number of coefficients:	1			
Window size:	125			
Step size:	125			
Full sample estimation results:				
Dependent Variable: D_LOML				
Method: Least Squares				
Date: 09/06/16 Time: 16:28				
Sample (adjusted): 1/05/2010 12/31/2015				
Included observations: 1499 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
D_LBVT	0.460128	0.024493	18.78626	0.0000
R-squared	0.189110	Mean dependent var		0.000677
Adjusted R-squared	0.189110	S.D. dependent var		0.015400
S.E. of regression	0.013868	Akaike info criterion		-5.717855
Sum squared resid	0.288081	Schwarz criterion		-5.714311
Log likelihood	4286.533	Hannan-Quinn criter.		-5.716535
Durbin-Watson stat	2.121056			

Appendix A1.6 Rolling regression of the financial sector

Roll:				
Roll type: Fixed window				
Manual specification: D_LREI D_LBVT				
Estimation command: ROLL(F,WINDOW=125,STEP=125) @ D_LREI D_LBVT				
Number of subsamples:	11			
Number of coefficients:	1			
Window size:	125			
Step size:	125			
Full sample estimation results:				
Dependent Variable: D_LREI				
Method: Least Squares				
Date: 09/06/16 Time: 16:30				
Sample (adjusted): 1/05/2010 12/31/2015				
Included observations: 1499 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
D_LBVT	0.229086	0.019774	11.58520	0.0000
R-squared	0.079305	Mean dependent var		0.000659
Adjusted R-squared	0.079305	S.D. dependent var		0.011668
S.E. of regression	0.011196	Akaike info criterion		-6.145875
Sum squared resid	0.187771	Schwarz criterion		-6.142331
Log likelihood	4607.334	Hannan-Quinn criter.		-6.144555
Durbin-Watson stat	2.242370			

Appendix A1.7 Rolling regression of the financial sector

Roll:				
Roll type: Fixed window				
Manual specification: D_LRMH D_LBVT				
Estimation command: ROLL(F,WINDOW=125,STEP=125) @ D_LRMH D_LBVT				
Number of subsamples:	11			
Number of coefficients:	1			
Window size:	125			
Step size:	125			
Full sample estimation results:				
Dependent Variable: D_LRMH				
Method: Least Squares				
Date: 09/06/16 Time: 16:33				
Sample (adjusted): 1/05/2010 12/31/2015				
Included observations: 1499 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
D_LBVT	0.662796	0.030418	21.78934	0.0000
R-squared	0.240316	Mean dependent var		0.000423
Adjusted R-squared	0.240316	S.D. dependent var		0.019760
S.E. of regression	0.017223	Akaike info criterion		-5.284521
Sum squared resid	0.444334	Schwarz criterion		-5.280977
Log likelihood	3961.748	Hannan-Quinn criter.		-5.283201
Durbin-Watson stat	2.239583			

Appendix A1.8 Rolling regression of the financial sector

Roll:				
Roll type: Fixed window				
Manual specification: D_LSLM D_LBVT				
Estimation command: ROLL(F,WINDOW=125,STEP=125) @ D_LSLM D_LBVT				
Number of subsamples:	11			
Number of coefficients:	1			
Window size:	125			
Step size:	125			
Full sample estimation results:				
Dependent Variable: D_LSLM				
Method: Least Squares				
Date: 09/06/16 Time: 16:35				
Sample (adjusted): 1/05/2010 12/31/2015				
Included observations: 1499 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
D_LBVT	0.592942	0.023420	25.31819	0.0000
R-squared	0.298490	Mean dependent var		0.000651
Adjusted R-squared	0.298490	S.D. dependent var		0.015832
S.E. of regression	0.013260	Akaike info criterion		-5.807468
Sum squared resid	0.263388	Schwarz criterion		-5.803924
Log likelihood	4353.697	Hannan-Quinn criter.		-5.806147
Durbin-Watson stat	2.081466			

Appendix A2.1 Cointegration analysis of the financial sector

Dependent Variable: NED				
Method: Least Squares				
Date: 05/26/16 Time: 16:13				
Sample: 4/01/2010 12/30/2015				
Included observations: 1500				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
FSR	3.319055	0.026364	125.8923	0.0000
C	7770.985	89.20638	87.11244	0.0000
R-squared	0.913644	Mean dependent var		18279.53
Adjusted R-squared	0.913587	S.D. dependent var		4145.726
S.E. of regression	1218.683	Akaike info criterion		17.05026
Sum squared resid	2.22E+09	Schwarz criterion		17.05735
Log likelihood	-12785.70	Hannan-Quinn criter.		17.05290
F-statistic	15848.88	Durbin-Watson stat		0.029878
Prob(F-statistic)	0.000000			

Appendix A2.2 Cointegration analysis of the financial sector

Dependent Variable: RMH				
Method: Least Squares				
Date: 05/26/16 Time: 16:21				
Sample: 4/01/2010 12/30/2015				
Included observations: 1500				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
FSR	1.112573	0.006496	171.2662	0.0000
C	788.7891	21.98050	35.88585	0.0000
R-squared	0.951411	Mean dependent var		4311.336
Adjusted R-squared	0.951379	S.D. dependent var		1361.819
S.E. of regression	300.2841	Akaike info criterion		14.24867
Sum squared resid	1.35E+08	Schwarz criterion		14.25575
Log likelihood	-10684.50	Hannan-Quinn criter.		14.25131
F-statistic	29332.13	Durbin-Watson stat		0.026309
Prob(F-statistic)	0.000000			

Appendix A2.3 Cointegration analysis of the financial sector

Dependent Variable: RMH				
Method: Least Squares				
Date: 05/26/16 Time: 16:24				
Sample: 4/01/2010 12/30/2015				
Included observations: 1500				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
INL	0.565966	0.004693	120.6019	0.0000
C	331.1975	34.70840	9.542286	0.0000
R-squared	0.906625	Mean dependent var		4311.336
Adjusted R-squared	0.906563	S.D. dependent var		1361.819
S.E. of regression	416.2743	Akaike info criterion		14.90190
Sum squared resid	2.60E+08	Schwarz criterion		14.90898
Log likelihood	-11174.42	Hannan-Quinn criter.		14.90454
F-statistic	14544.83	Durbin-Watson stat		0.032903
Prob(F-statistic)	0.000000			

Appendix A3.1 ECM analysis of the financial sector

Null Hypothesis: ECM_0 has a unit root				
Exogenous: Constant				
Lag Length: 1 (Automatic - based on SIC, maxlag=23)				
			t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic			-2.452531	0.1276
Test critical values:	1% level		-3.434508	
	5% level		-2.863264	
	10% level		-2.567736	
*MacKinnon (1996) one-sided p-values.				
Augmented Dickey-Fuller Test Equation				
Dependent Variable: D(ECM_0)				
Method: Least Squares				
Date: 05/26/16 Time: 16:14				
Sample (adjusted): 4/05/2010 12/30/2015				
Included observations: 1498 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
ECM_0(-1)	-0.010874	0.004434	-2.452531	0.0143
D(ECM_0(-1))	-0.155459	0.025588	-6.075466	0.0000
C	-1.181856	5.365405	-0.220273	0.8257
R-squared	0.029975	Mean dependent var		-1.056141
Adjusted R-squared	0.028678	S.D. dependent var		210.7026
S.E. of regression	207.6594	Akaike info criterion		13.51168
Sum squared resid	64468006	Schwarz criterion		13.52231
Log likelihood	-10117.25	Hannan-Quinn criter.		13.51564
F-statistic	23.09897	Durbin-Watson stat		2.011897
Prob(F-statistic)	0.000000			

Appendix A3.2 ECM analysis of the financial sector

Null Hypothesis: ECM_1 has a unit root				
Exogenous: Constant				
Lag Length: 1 (Automatic - based on SIC, maxlag=23)				
			t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic			-2.663830	0.0807
Test critical values:	1% level		-3.434508	
	5% level		-2.863264	
	10% level		-2.567736	
*MacKinnon (1996) one-sided p-values.				
Augmented Dickey-Fuller Test Equation				
Dependent Variable: D(ECM_1)				
Method: Least Squares				
Date: 05/26/16 Time: 16:22				
Sample (adjusted): 4/05/2010 12/30/2015				
Included observations: 1498 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
ECM_1(-1)	-0.011019	0.004137	-2.663830	0.0078
D(ECM_1(-1))	-0.166282	0.025532	-6.512745	0.0000
C	-0.047220	1.238005	-0.038142	0.9696
R-squared	0.034037	Mean dependent var		-0.049840
Adjusted R-squared	0.032744	S.D. dependent var		48.72001
S.E. of regression	47.91573	Akaike info criterion		10.57877
Sum squared resid	3432396.	Schwarz criterion		10.58940
Log likelihood	-7920.495	Hannan-Quinn criter.		10.58273
F-statistic	26.33878	Durbin-Watson stat		2.012180
Prob(F-statistic)	0.000000			

Appendix A3.3 ECM analysis of the financial sector

Null Hypothesis: ECM_2 has a unit root				
Exogenous: Constant				
Lag Length: 1 (Automatic - based on SIC, maxlag=23)				
			t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic			-2.702269	0.0738
Test critical values:	1% level		-3.434508	
	5% level		-2.863264	
	10% level		-2.567736	
*MacKinnon (1996) one-sided p-values.				
Augmented Dickey-Fuller Test Equation				
Dependent Variable: D(ECM_2)				
Method: Least Squares				
Date: 05/26/16 Time: 16:26				
Sample (adjusted): 4/05/2010 12/30/2015				
Included observations: 1498 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
ECM_2(-1)	-0.012639	0.004677	-2.702269	0.0070
D(ECM_2(-1))	-0.116879	0.025730	-4.542499	0.0000
C	-0.500758	1.933077	-0.259047	0.7956
R-squared	0.020125	Mean dependent var		-0.460679
Adjusted R-squared	0.018814	S.D. dependent var		75.53015
S.E. of regression	74.81626	Akaike info criterion		11.46995
Sum squared resid	8368222.	Schwarz criterion		11.48059
Log likelihood	-8587.991	Hannan-Quinn criter.		11.47391
F-statistic	15.35234	Durbin-Watson stat		2.000576
Prob(F-statistic)	0.000000			

Appendix B1.1 Rolling regression of the industrial sector

Roll:				
Roll type: Fixed window				
Manual specification: DL_NMK DL_MTN				
Estimation command: ROLL(F,WINDOW=125,STEP=125) @ DL_NMK				
DL_MTN				
Number of subsamples:	11			
Number of coefficients:	1			
Window size:	125			
Step size:	125			
Full sample estimation results:				
Dependent Variable: DL_NMK				
Method: Least Squares				
Date: 09/03/16 Time: 12:34				
Sample (adjusted): 1/05/2010 12/31/2015				
Included observations: 1499 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
DL_MTN	0.189513	0.023975	7.904620	0.0000
R-squared	0.039659	Mean dependent var		0.000336
Adjusted R-squared	0.039659	S.D. dependent var		0.016860
S.E. of regression	0.016522	Akaike info criterion		-5.367560
Sum squared resid	0.408927	Schwarz criterion		-5.364016
Log likelihood	4023.986	Hannan-Quinn criter.		-5.366240
Durbin-Watson stat	2.076020			

Appendix B1.2 Rolling regression of the industrial sector

Roll:				
Roll type: Fixed window				
Manual specification: DL_NTC DL_MND				
Estimation command: ROLL(F,WINDOW=125,STEP=125) @ DL_NTC DL_MND				
Number of subsamples:	11			
Number of coefficients:	1			
Window size:	125			
Step size:	125			
Full sample estimation results:				
Dependent Variable: DL_NTC				
Method: Least Squares				
Date: 09/03/16 Time: 12:38				
Sample (adjusted): 1/05/2010 12/31/2015				
Included observations: 1499 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
DL_MND	0.210553	0.019955	10.55119	0.0000
R-squared	0.067559	Mean dependent var		0.000608
Adjusted R-squared	0.067559	S.D. dependent var		0.014586
S.E. of regression	0.014084	Akaike info criterion		-5.686843
Sum squared resid	0.297155	Schwarz criterion		-5.683299
Log likelihood	4263.289	Hannan-Quinn criter.		-5.685522
Durbin-Watson stat	2.253208			

Appendix B3 Cointegration analysis of the industrial sector

Dependent Variable: L_NMK				
Method: Least Squares				
Date: 08/30/16 Time: 15:43				
Sample: 1/04/2010 12/31/2015				
Included observations: 1500				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
L_MTN	1.133463	0.009387	120.7437	0.0000
C	-3.067162	0.091032	-33.69336	0.0000
R-squared	0.906824	Mean dependent var		7.920902
Adjusted R-squared	0.906761	S.D. dependent var		0.288738
S.E. of regression	0.088166	Akaike info criterion		-2.017855
Sum squared resid	11.64435	Schwarz criterion		-2.010771
Log likelihood	1515.392	Hannan-Quinn criter.		-2.015216
F-statistic	14579.05	Durbin-Watson stat		0.071459
Prob(F-statistic)	0.000000			

Appendix B4.1 ECM analysis of the industrial sector

Null Hypothesis: ECM has a unit root				
Exogenous: Constant				
Lag Length: 1 (Automatic - based on SIC, maxlag=23)				
			t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic			-4.873448	0.0000
Test critical values:	1% level		-3.434508	
	5% level		-2.863264	
	10% level		-2.567736	
*MacKinnon (1996) one-sided p-values.				
Augmented Dickey-Fuller Test Equation				
Dependent Variable: D(ECM)				
Method: Least Squares				
Date: 08/30/16 Time: 15:18				
Sample (adjusted): 1/06/2010 12/31/2015				
Included observations: 1498 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
ECM(-1)	-0.033586	0.006892	-4.873448	0.0000
D(ECM(-1))	-0.077374	0.025732	-3.006909	0.0027
C	0.000206	0.000601	0.343471	0.7313
R-squared	0.024244	Mean dependent var		0.000189
Adjusted R-squared	0.022939	S.D. dependent var		0.023534
S.E. of regression	0.023263	Akaike info criterion		-4.681922
Sum squared resid	0.809031	Schwarz criterion		-4.671284
Log likelihood	3509.760	Hannan-Quinn criter.		-4.677959
F-statistic	18.57302	Durbin-Watson stat		1.998761
Prob(F-statistic)	0.000000			

Appendix B4.2 ECM analysis of the industrial sector

Dependent Variable: DL_NMK				
Method: Least Squares				
Date: 08/31/16 Time: 11:57				
Sample (adjusted): 1/05/2010 12/31/2015				
Included observations: 1499 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
DL_MTN	0.198251	0.024049	8.243491	0.0000
ECM(-1)	-0.016045	0.004860	-3.301365	0.0010
C	0.000315	0.000425	0.741574	0.4585
R-squared	0.046957	Mean dependent var		0.000336
Adjusted R-squared	0.045683	S.D. dependent var		0.016860
S.E. of regression	0.016470	Akaike info criterion		-5.372520
Sum squared resid	0.405820	Schwarz criterion		-5.361888
Log likelihood	4029.704	Hannan-Quinn criter.		-5.368559
F-statistic	36.85452	Durbin-Watson stat		2.057455
Prob(F-statistic)	0.000000			

Appendix B4.3 ECM analysis of the industrial sector

Dependent Variable: DL_MTN				
Method: Least Squares				
Date: 08/31/16 Time: 13:34				
Sample (adjusted): 1/05/2010 12/31/2015				
Included observations: 1499 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
DL_NMK	0.219170	0.026587	8.243491	0.0000
ECM(-1)	0.025032	0.005088	4.920267	0.0000
C	2.68E-05	0.000447	0.059992	0.9522
R-squared	0.055301	Mean dependent var		9.85E-05
Adjusted R-squared	0.054038	S.D. dependent var		0.017805
S.E. of regression	0.017317	Akaike info criterion		-5.272207
Sum squared resid	0.448641	Schwarz criterion		-5.261575
Log likelihood	3954.519	Hannan-Quinn criter.		-5.268246
F-statistic	43.78692	Durbin-Watson stat		2.156971
Prob(F-statistic)	0.000000			

Appendix C1.1 Rolling regression analysis of the mining sector

Roll:				
Roll type: Fixed window				
Manual specification: D_LAMS D_LACL				
Estimation command: ROLL(F,WINDOW=125,STEP=125) @ D_LAMS				
D_LACL				
Number of subsamples:	11			
Number of coefficients:	1			
Window size:	125			
Step size:	125			
Full sample estimation results:				
Dependent Variable: D_LAMS				
Method: Least Squares				
Date: 09/03/16 Time: 11:52				
Sample (adjusted): 1/05/2010 12/31/2015				
Included observations: 1499 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
D_LACL	0.170702	0.018024	9.470837	0.0000
R-squared	0.054747	Mean dependent var		-0.000980
Adjusted R-squared	0.054747	S.D. dependent var		0.022771
S.E. of regression	0.022139	Akaike info criterion		-4.782279
Sum squared resid	0.734227	Schwarz criterion		-4.778735
Log likelihood	3585.318	Hannan-Quinn criter.		-4.780959
Durbin-Watson stat	1.962793			

Appendix C1.2 Rolling regression analysis of the mining sector

Roll:				
Roll type: Fixed window				
Manual specification: D_LIMP D_LACL				
Estimation command: ROLL(F,WINDOW=125,STEP=125) @ D_LIMP D_LACL				
Number of subsamples:	11			
Number of coefficients:	1			
Window size:	125			
Step size:	125			
Full sample estimation results:				
Dependent Variable: D_LIMP				
Method: Least Squares				
Date: 09/03/16 Time: 11:55				
Sample (adjusted): 1/05/2010 12/31/2015				
Included observations: 1499 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
D_LACL	0.227663	0.018012	12.63926	0.0000
R-squared	0.092971	Mean dependent var		-0.001423
Adjusted R-squared	0.092971	S.D. dependent var		0.023231
S.E. of regression	0.022125	Akaike info criterion		-4.783570
Sum squared resid	0.733280	Schwarz criterion		-4.780026
Log likelihood	3586.286	Hannan-Quinn criter.		-4.782250
Durbin-Watson stat	1.987527			

Appendix C1.3 Rolling regression analysis of the mining sector

Roll:				
Roll type: Fixed window				
Manual specification: D_LIMP D_LAGL				
Estimation command: ROLL(F,WINDOW=125,STEP=125) @ D_LIMP D_LAGL				
Number of subsamples:	11			
Number of coefficients:	1			
Window size:	125			
Step size:	125			
Full sample estimation results:				
Dependent Variable: D_LIMP				
Method: Least Squares				
Date: 09/03/16 Time: 11:57				
Sample (adjusted): 1/05/2010 12/31/2015				
Included observations: 1499 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
D_LAGL	0.614845	0.023344	26.33870	0.0000
R-squared	0.313953	Mean dependent var		-0.001423
Adjusted R-squared	0.313953	S.D. dependent var		0.023231
S.E. of regression	0.019242	Akaike info criterion		-5.062798
Sum squared resid	0.554629	Schwarz criterion		-5.059254
Log likelihood	3795.567	Hannan-Quinn criter.		-5.061478
Durbin-Watson stat	2.105599			

Appendix C1.4 Rolling regression analysis of the mining sector

Roll:				
Roll type: Fixed window				
Manual specification: D_LIMP D_LAMS				
Estimation command: ROLL(F,WINDOW=125,STEP=125) @ D_LIMP D_LAMS				
Number of subsamples:	11			
Number of coefficients:	1			
Window size:	125			
Step size:	125			
Full sample estimation results:				
Dependent Variable: D_LIMP				
Method: Least Squares				
Date: 09/03/16 Time: 12:00				
Sample (adjusted): 1/05/2010 12/31/2015				
Included observations: 1499 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
D_LAMS	0.699198	0.019229	36.36143	0.0000
R-squared	0.466828	Mean dependent var		-0.001423
Adjusted R-squared	0.466828	S.D. dependent var		0.023231
S.E. of regression	0.016963	Akaike info criterion		-5.314899
Sum squared resid	0.431039	Schwarz criterion		-5.311355
Log likelihood	3984.517	Hannan-Quinn criter.		-5.313579
Durbin-Watson stat	2.229591			

Appendix C1.5 Rolling regression analysis of the mining sector

Roll:				
Roll type: Fixed window				
Manual specification: D_LGFI D_LANG				
Estimation command: ROLL(F,WINDOW=125,STEP=125) @ D_LGFI D_LANG				
Number of subsamples:	11			
Number of coefficients:	1			
Window size:	125			
Step size:	125			
Full sample estimation results:				
Dependent Variable: D_LGFI				
Method: Least Squares				
Date: 09/03/16 Time: 12:03				
Sample (adjusted): 1/05/2010 12/31/2015				
Included observations: 1499 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
D_LANG	0.770046	0.018358	41.94566	0.0000
R-squared	0.539914	Mean dependent var		-0.000571
Adjusted R-squared	0.539914	S.D. dependent var		0.026395
S.E. of regression	0.017904	Akaike info criterion		-5.206952
Sum squared resid	0.480173	Schwarz criterion		-5.203408
Log likelihood	3903.610	Hannan-Quinn criter.		-5.205631
Durbin-Watson stat	2.017295			

Appendix C2 Cointegration analysis of the mining sector

Dependent Variable: LAMS				
Method: Least Squares				
Date: 09/05/16 Time: 14:14				
Sample: 1/04/2010 12/31/2015				
Included observations: 1500				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
LACL	0.471805	0.004415	106.8566	0.0000
C	6.809428	0.036854	184.7682	0.0000
R-squared	0.884023	Mean dependent var		10.73479
Adjusted R-squared	0.883945	S.D. dependent var		0.336449
S.E. of regression	0.114618	Akaike info criterion		-1.493100
Sum squared resid	19.67949	Schwarz criterion		-1.486016
Log likelihood	1121.825	Hannan-Quinn criter.		-1.490461
F-statistic	11418.33	Durbin-Watson stat		0.044260
Prob(F-statistic)	0.000000			

Appendix C3.1 ECM analysis of the mining sector

Null Hypothesis: ECM1 has a unit root				
Exogenous: Constant				
Lag Length: 0 (Automatic - based on SIC, maxlag=23)				
			t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic			-4.076035	0.0011
Test critical values:	1% level		-3.434505	
	5% level		-2.863262	
	10% level		-2.567735	
*MacKinnon (1996) one-sided p-values.				
Augmented Dickey-Fuller Test Equation				
Dependent Variable: D(ECM1)				
Method: Least Squares				
Date: 09/05/16 Time: 14:22				
Sample (adjusted): 1/05/2010 12/31/2015				
Included observations: 1499 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
ECM1(-1)	-0.022052	0.005410	-4.076035	0.0000
C	8.16E-06	0.000620	0.013177	0.9895
R-squared	0.010976	Mean dependent var		1.02E-05
Adjusted R-squared	0.010316	S.D. dependent var		0.024113
S.E. of regression	0.023989	Akaike info criterion		-4.621141
Sum squared resid	0.861454	Schwarz criterion		-4.614053
Log likelihood	3465.545	Hannan-Quinn criter.		-4.618501
F-statistic	16.61406	Durbin-Watson stat		1.967698
Prob(F-statistic)	0.000048			

Appendix C3.2 ECM analysis of the mining sector

Null Hypothesis: ECM2 has a unit root				
Exogenous: Constant				
Lag Length: 0 (Automatic - based on SIC, maxlag=23)				
			t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic			-3.275765	0.0162
Test critical values:	1% level		-3.434505	
	5% level		-2.863262	
	10% level		-2.567735	
*MacKinnon (1996) one-sided p-values.				
Augmented Dickey-Fuller Test Equation				
Dependent Variable: D(ECM2)				
Method: Least Squares				
Date: 09/05/16 Time: 14:21				
Sample (adjusted): 1/05/2010 12/31/2015				
Included observations: 1499 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
ECM2(-1)	-0.014200	0.004335	-3.275765	0.0011
C	1.06E-05	0.000680	0.015552	0.9876
R-squared	0.007117	Mean dependent var		9.83E-06
Adjusted R-squared	0.006454	S.D. dependent var		0.026428
S.E. of regression	0.026343	Akaike info criterion		-4.433930
Sum squared resid	1.038812	Schwarz criterion		-4.426842
Log likelihood	3325.231	Hannan-Quinn criter.		-4.431290
F-statistic	10.73064	Durbin-Watson stat		1.954395
Prob(F-statistic)	0.001078			

Appendix C3.3 ECM analysis of the mining sector

Null Hypothesis: ECM3 has a unit root				
Exogenous: Constant				
Lag Length: 1 (Automatic - based on SIC, maxlag=23)				
			t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic			-2.266360	0.1832
Test critical values:	1% level		-3.434508	
	5% level		-2.863264	
	10% level		-2.567736	
*MacKinnon (1996) one-sided p-values.				
Augmented Dickey-Fuller Test Equation				
Dependent Variable: D(ECM3)				
Method: Least Squares				
Date: 09/05/16 Time: 14:24				
Sample (adjusted): 1/06/2010 12/31/2015				
Included observations: 1498 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
ECM3(-1)	-0.008999	0.003971	-2.266360	0.0236
D(ECM3(-1))	-0.080212	0.025822	-3.106270	0.0019
C	0.000210	0.000720	0.291918	0.7704
R-squared	0.010698	Mean dependent var		0.000196
Adjusted R-squared	0.009375	S.D. dependent var		0.028000
S.E. of regression	0.027868	Akaike info criterion		-4.320640
Sum squared resid	1.161095	Schwarz criterion		-4.310002
Log likelihood	3239.160	Hannan-Quinn criter.		-4.316677
F-statistic	8.083271	Durbin-Watson stat		1.998104
Prob(F-statistic)	0.000322			

Appendix C3.4 ECM analysis of the mining sector

Null Hypothesis: ECM4 has a unit root				
Exogenous: Constant				
Lag Length: 1 (Automatic - based on SIC, maxlag=23)				
			t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic			-1.571539	0.4969
Test critical values:	1% level		-3.434508	
	5% level		-2.863264	
	10% level		-2.567736	
*MacKinnon (1996) one-sided p-values.				
Augmented Dickey-Fuller Test Equation				
Dependent Variable: D(ECM4)				
Method: Least Squares				
Date: 09/05/16 Time: 14:26				
Sample (adjusted): 1/06/2010 12/31/2015				
Included observations: 1498 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
ECM4(-1)	-0.004666	0.002969	-1.571539	0.1163
D(ECM4(-1))	-0.115826	0.025734	-4.500907	0.0000
C	-0.000161	0.000567	-0.284752	0.7759
R-squared	0.015623	Mean dependent var		-0.000152
Adjusted R-squared	0.014306	S.D. dependent var		0.022101
S.E. of regression	0.021942	Akaike info criterion		-4.798794
Sum squared resid	0.719794	Schwarz criterion		-4.788156
Log likelihood	3597.297	Hannan-Quinn criter.		-4.794831
F-statistic	11.86330	Durbin-Watson stat		2.005318
Prob(F-statistic)	0.000008			

Appendix D1.1 Rolling regression analysis of the retail sector

Roll:				
Roll type: Fixed window				
Manual specification: D_LSAB D_LAPN				
Estimation command: ROLL(F,WINDOW=125,STEP=125) @ D_LSAB D_LAPN				
Number of subsamples:	11			
Number of coefficients:	1			
Window size:	125			
Step size:	125			
Full sample estimation results:				
Dependent Variable: D_LSAB				
Method: Least Squares				
Date: 09/08/16 Time: 14:38				
Sample (adjusted): 1/05/2010 12/31/2015				
Included observations: 1499 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
D_LAPN	0.241224	0.022295	10.81954	0.0000
R-squared	0.068269	Mean dependent var		0.000971
Adjusted R-squared	0.068269	S.D. dependent var		0.014410
S.E. of regression	0.013910	Akaike info criterion		-5.711798
Sum squared resid	0.289831	Schwarz criterion		-5.708254
Log likelihood	4281.993	Hannan-Quinn criter.		-5.710478
Durbin-Watson stat	2.086161			

Appendix D1.2 Rolling regression analysis of the retail sector

Roll:				
Roll type: Fixed window				
Manual specification: D_LSAB D_LBTI				
Estimation command: ROLL(F,WINDOW=125,STEP=125) @ D_LSAB D_LBTI				
Number of subsamples:	11			
Number of coefficients:	1			
Window size:	125			
Step size:	125			
Full sample estimation results:				
Dependent Variable: D_LSAB				
Method: Least Squares				
Date: 09/08/16 Time: 14:42				
Sample (adjusted): 1/05/2010 12/31/2015				
Included observations: 1499 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
D_LBTI	0.698920	0.028326	24.67441	0.0000
R-squared	0.285749	Mean dependent var		0.000971
Adjusted R-squared	0.285749	S.D. dependent var		0.014410
S.E. of regression	0.012179	Akaike info criterion		-5.977608
Sum squared resid	0.222180	Schwarz criterion		-5.974064
Log likelihood	4481.217	Hannan-Quinn criter.		-5.976287
Durbin-Watson stat	2.120134			

Appendix D1.3 Rolling regression analysis of the retail sector

Roll:				
Roll type: Fixed window				
Manual specification: D_LWHL D_LBTI				
Estimation command: ROLL(F,WINDOW=125,STEP=125) @ D_LWHL D_LBTI				
Number of subsamples:	11			
Number of coefficients:	1			
Window size:	125			
Step size:	125			
Full sample estimation results:				
Dependent Variable: D_LWHL				
Method: Least Squares				
Date: 09/08/16 Time: 14:45				
Sample (adjusted): 1/05/2010 12/31/2015				
Included observations: 1499 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
D_LBTI	0.191704	0.040028	4.789219	0.0000
R-squared	0.010737	Mean dependent var		0.001149
Adjusted R-squared	0.010737	S.D. dependent var		0.017303
S.E. of regression	0.017210	Akaike info criterion		-5.285979
Sum squared resid	0.443686	Schwarz criterion		-5.282435
Log likelihood	3962.842	Hannan-Quinn criter.		-5.284659
Durbin-Watson stat	1.939831			

Appendix D1.4 Rolling regression analysis of the retail sector

Roll:				
Roll type: Fixed window				
Manual specification: D_LWHL D_LSAB				
Estimation command: ROLL(F,WINDOW=125,STEP=125) @ D_LWHL D_LSAB				
Number of subsamples:	11			
Number of coefficients:	1			
Window size:	125			
Step size:	125			
Full sample estimation results:				
Dependent Variable: D_LWHL				
Method: Least Squares				
Date: 09/08/16 Time: 14:48				
Sample (adjusted): 1/05/2010 12/31/2015				
Included observations: 1499 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
D_LSAB	0.229444	0.030450	7.535020	0.0000
R-squared	0.032269	Mean dependent var		0.001149
Adjusted R-squared	0.032269	S.D. dependent var		0.017303
S.E. of regression	0.017022	Akaike info criterion		-5.307985
Sum squared resid	0.434030	Schwarz criterion		-5.304441
Log likelihood	3979.335	Hannan-Quinn criter.		-5.306665
Durbin-Watson stat	1.930874			

Appendix D1.5 Rolling regression analysis of the retail sector

Roll:				
Roll type: Fixed window				
Manual specification: D_LWHL D_LTSG				
Estimation command: ROLL(F,WINDOW=125,STEP=125) @ D_LWHL D_LTSG				
Number of subsamples:	11			
Number of coefficients:	1			
Window size:	125			
Step size:	125			
Full sample estimation results:				
Dependent Variable: D_LWHL				
Method: Least Squares				
Date: 09/08/16 Time: 14:52				
Sample (adjusted): 1/05/2010 12/31/2015				
Included observations: 1499 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
D_LTSG	0.549047	0.030426	18.04506	0.0000
R-squared	0.174936	Mean dependent var		0.001149
Adjusted R-squared	0.174936	S.D. dependent var		0.017303
S.E. of regression	0.015717	Akaike info criterion		-5.467479
Sum squared resid	0.370043	Schwarz criterion		-5.463935
Log likelihood	4098.876	Hannan-Quinn criter.		-5.466159
Durbin-Watson stat	2.042049			

Appendix D2 Cointegration analysis of the retail sector

Dependent Variable: LBTI				
Method: Least Squares				
Date: 09/08/16 Time: 15:15				
Sample: 1/04/2010 12/31/2015				
Included observations: 1500				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
LAPN	0.594277	0.004960	119.8105	0.0000
C	4.914523	0.048303	101.7445	0.0000
R-squared	0.905504	Mean dependent var		10.69081
Adjusted R-squared	0.905441	S.D. dependent var		0.372754
S.E. of regression	0.114623	Akaike info criterion		-1.492999
Sum squared resid	19.68148	Schwarz criterion		-1.485914
Log likelihood	1121.749	Hannan-Quinn criter.		-1.490359
F-statistic	14354.56	Durbin-Watson stat		0.013225
Prob(F-statistic)	0.000000			

Appendix D3.1 ECM analysis of the retail sector

Null Hypothesis: ECM1 has a unit root				
Exogenous: Constant				
Lag Length: 0 (Automatic - based on SIC, maxlag=23)				
			t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic			-1.489261	0.5390
Test critical values:	1% level		-3.434505	
	5% level		-2.863262	
	10% level		-2.567735	
*MacKinnon (1996) one-sided p-values.				
Augmented Dickey-Fuller Test Equation				
Dependent Variable: D(ECM1)				
Method: Least Squares				
Date: 09/08/16 Time: 15:17				
Sample (adjusted): 1/05/2010 12/31/2015				
Included observations: 1499 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
ECM1(-1)	-0.004434	0.002977	-1.489261	0.1366
C	0.000292	0.000340	0.857115	0.3915
R-squared	0.001479	Mean dependent var		0.000293
Adjusted R-squared	0.000812	S.D. dependent var		0.013178
S.E. of regression	0.013173	Akaike info criterion		-5.819948
Sum squared resid	0.259775	Schwarz criterion		-5.812860
Log likelihood	4364.051	Hannan-Quinn criter.		-5.817307
F-statistic	2.217898	Durbin-Watson stat		1.991557
Prob(F-statistic)	0.136629			

Appendix D3.2 ECM analysis of the retail sector

Null Hypothesis: ECM2 has a unit root				
Exogenous: Constant				
Lag Length: 0 (Automatic - based on SIC, maxlag=23)				
			t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic			-1.309215	0.6273
Test critical values:	1% level		-3.434505	
	5% level		-2.863262	
	10% level		-2.567735	
*MacKinnon (1996) one-sided p-values.				
Augmented Dickey-Fuller Test Equation				
Dependent Variable: D(ECM2)				
Method: Least Squares				
Date: 09/08/16 Time: 15:22				
Sample (adjusted): 1/05/2010 12/31/2015				
Included observations: 1499 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
ECM2(-1)	-0.004820	0.003682	-1.309215	0.1907
C	0.000301	0.000406	0.741365	0.4586
R-squared	0.001144	Mean dependent var		0.000302
Adjusted R-squared	0.000476	S.D. dependent var		0.015723
S.E. of regression	0.015719	Akaike info criterion		-5.466497
Sum squared resid	0.369912	Schwarz criterion		-5.459409
Log likelihood	4099.140	Hannan-Quinn criter.		-5.463856
F-statistic	1.714045	Durbin-Watson stat		2.019389
Prob(F-statistic)	0.190662			

Appendix D3.3 ECM analysis of the retail sector

Null Hypothesis: ECM3 has a unit root				
Exogenous: Constant				
Lag Length: 0 (Automatic - based on SIC, maxlag=23)				
			t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic			-3.857558	0.0024
Test critical values:	1% level		-3.434505	
	5% level		-2.863262	
	10% level		-2.567735	
*MacKinnon (1996) one-sided p-values.				
Augmented Dickey-Fuller Test Equation				
Dependent Variable: D(ECM3)				
Method: Least Squares				
Date: 09/08/16 Time: 15:21				
Sample (adjusted): 1/05/2010 12/31/2015				
Included observations: 1499 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
ECM3(-1)	-0.019394	0.005028	-3.857558	0.0001
C	-1.41E-05	0.000337	-0.041644	0.9668
R-squared	0.009843	Mean dependent var		-1.29E-05
Adjusted R-squared	0.009181	S.D. dependent var		0.013127
S.E. of regression	0.013067	Akaike info criterion		-5.836183
Sum squared resid	0.255591	Schwarz criterion		-5.829095
Log likelihood	4376.219	Hannan-Quinn criter.		-5.833542
F-statistic	14.88076	Durbin-Watson stat		2.078282
Prob(F-statistic)	0.000119			

Appendix D3.4 ECM analysis of the retail sector

Null Hypothesis: ECM4 has a unit root Exogenous: Constant Lag Length: 0 (Automatic - based on SIC, maxlag=23)				
			t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic			-3.680438	0.0045
Test critical values:	1% level		-3.434505	
	5% level		-2.863262	
	10% level		-2.567735	
*MacKinnon (1996) one-sided p-values.				
Augmented Dickey-Fuller Test Equation				
Dependent Variable: D(ECM4)				
Method: Least Squares				
Date: 09/08/16 Time: 15:23				
Sample (adjusted): 1/05/2010 12/31/2015				
Included observations: 1499 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
ECM4(-1)	-0.017679	0.004804	-3.680438	0.0002
C	1.58E-05	0.000547	0.028907	0.9769
R-squared	0.008967	Mean dependent var		1.30E-05
Adjusted R-squared	0.008305	S.D. dependent var		0.021265
S.E. of regression	0.021177	Akaike info criterion		-4.870510
Sum squared resid	0.671325	Schwarz criterion		-4.863421
Log likelihood	3652.447	Hannan-Quinn criter.		-4.867869
F-statistic	13.54562	Durbin-Watson stat		1.993282
Prob(F-statistic)	0.000241			

Appendix D3.5 ECM analysis of the retail sector

Null Hypothesis: ECM5 has a unit root				
Exogenous: Constant				
Lag Length: 0 (Automatic - based on SIC, maxlag=23)				
			t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic			-3.257157	0.0171
Test critical values:	1% level		-3.434505	
	5% level		-2.863262	
	10% level		-2.567735	
*MacKinnon (1996) one-sided p-values.				
Augmented Dickey-Fuller Test Equation				
Dependent Variable: D(ECM5)				
Method: Least Squares				
Date: 09/08/16 Time: 15:25				
Sample (adjusted): 1/05/2010 12/31/2015				
Included observations: 1499 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
ECM5(-1)	-0.012917	0.003966	-3.257157	0.0012
C	5.90E-05	0.000550	0.107183	0.9147
R-squared	0.007037	Mean dependent var		5.62E-05
Adjusted R-squared	0.006374	S.D. dependent var		0.021381
S.E. of regression	0.021313	Akaike info criterion		-4.857684
Sum squared resid	0.679990	Schwarz criterion		-4.850596
Log likelihood	3642.834	Hannan-Quinn criter.		-4.855044
F-statistic	10.60907	Durbin-Watson stat		1.999538
Prob(F-statistic)	0.001151			

Appendix D3.6 ECM analysis of the retail sector

Null Hypothesis: ECM6 has a unit root				
Exogenous: Constant				
Lag Length: 1 (Automatic - based on SIC, maxlag=23)				
			t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic			-1.875642	0.3441
Test critical values:	1% level		-3.434508	
	5% level		-2.863264	
	10% level		-2.567736	
*MacKinnon (1996) one-sided p-values.				
Augmented Dickey-Fuller Test Equation				
Dependent Variable: D(ECM6)				
Method: Least Squares				
Date: 09/08/16 Time: 15:27				
Sample (adjusted): 1/06/2010 12/31/2015				
Included observations: 1498 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
ECM6(-1)	-0.005069	0.002702	-1.875642	0.0609
D(ECM6(-1))	-0.134445	0.025621	-5.247506	0.0000
C	7.21E-05	0.000567	0.127075	0.8989
R-squared	0.021045	Mean dependent var		6.32E-05
Adjusted R-squared	0.019735	S.D. dependent var		0.022167
S.E. of regression	0.021947	Akaike info criterion		-4.798368
Sum squared resid	0.720100	Schwarz criterion		-4.787730
Log likelihood	3596.978	Hannan-Quinn criter.		-4.794405
F-statistic	16.06911	Durbin-Watson stat		2.010651
Prob(F-statistic)	0.000000			