A Trading Strategy Based on the Lead-Lag Relationship of Spot and Futures Prices of the S&P 500

FE8827 Quantitative Trading Strategies
2010/11 Mini-Term 5
Nanyang Technological University

Submitted By:
Thursten Cheok Yong Jin - G0900101J
Ng Kok Keong – G0901861C
Kanika Jain – G0900518E

Contents

1. Introduction
2. The Theoretical Relationship between Spot and Futures Markets
3. Data Handling
4. Econometric Modeling
5. Formulating a Trading Strategy
6. Conclusion

1) Introduction

...

Introduction

- In theory the spot and futures prices of an asset (here, the S&P 500 Index) are mathematically related such that the returns are perfectly contemporaneously correlated.

- In practice, this correlation is often imperfect.

- This project aims to model the temporal relationship between the spot and futures prices of the S&P 500 and formulate a trading strategy based on this relationship.

2) The Theoretical Relationship between Spot and Futures Markets

• • •

Spot-Futures Relationship

- The theoretical spot-futures relationship is

$$F_t = S_t e^{(r-d)(T-t)} \qquad (1)$$

where
- F_t is the S&P 500 futures contract price quoted at time t
- S_t is the S&P 500 index level quoted at time t
- r is the continuously compounded risk-free rate
- d is the continuously compounded dividend yield
- T is the maturity date of the futures contract

- Under market efficiency and frictionless trading, the the spot and futures prices should be perfectly contemporaneously correlated according to Equation (1), such that neither market leads the other.
- In reality however, changes in the futures price often lead those in the spot price.

3) Data Handling

• • •

i. Data Sources
ii. Data Handling Steps

3) Data Handling

• • •

i. Data Sources
ii. Data Handling Steps

Data Handling

i. Data Sources

- Sample E-mini S&P 500 Futures tick-by-tick transaction data is downloaded from CQG Data Factory website
  - Data period from July 2007 to October 2007
  - Website: https://www.cqgdatafactory.com/?page=orderSample
- SPDR S&P 500 ETF (Symbol: SPY) tick-by-tick transaction data is downloaded from Wharton Research Data Services (WRDS) database through the NTU Library website
  - Data period from July 2007 to October 2007

3) Data Handling

• • •

i. Data Sources

ii. Data Handling Steps

Data Handling

ii. Data Handling Steps

- Step 1: Upload the tick-by-tick transaction data into 2 tables in an Access database, namely S&P500EminiFut and SPY.
- Step 2: Create a new column in both tables named TradeDT to record the 10-minute timestamp of the record in this format: "YYYYMMDDHHm", where "m" stands for the number of 10-minute of the hour.
- Step 3: Group the records by the TradeDT column and find the average price of each 10 minute using the following sql query:
  - SELECT TradeDT, avg(Price) FROM SP500EminiFut GROUP BY TradeDT
  - SELECT TradeDT, avg(Price) FROM SPY GROUP BY TradeDT

Data Handling

ii. Data Handling Steps

- Step 4: Place the 2 sets of data into one single Excel spreadsheet and match the records by the TradeDT values.
- Step 5: As the trading hours of NYSE is from 9:30am to 4:00pm, we remove all the records that are outside this trading hours.
- Step 6: If there are no transactions for Emini S&P 500 Futures or SPDR S&P 500 ETF, we assume that the price remains the same as the last available transaction.
- Step 7: 2 sets of data are now ready to be uploaded into EViews for analysis.

4) Econometric Modeling

• • •

i. Non-Stationarity Tests
ii. Estimating the Error Correction Model
iii. Estimating the Error Correction Model with Cost of Carry
iv. Estimating the Autoregressive Moving Average Model
v. Estimating the Vector Autoregressive Model
vi. Model Selection

4) Econometric Modeling

• • •

i. Non-Stationarity Tests
ii. Estimating the Error Correction Model
iii. Estimating the Error Correction Model with Cost of Carry
iv. Estimating the Autoregressive Moving Average Model
v. Estimating the Vector Autoregressive Model
vi. Model Selection

Econometric Modeling

i. Non-Stationarity Tests

- To test for non-stationarity, we apply the ADF and KPSS tests, consisting of the following hypotheses:

ADF Test	KPSS Test
H_0: There is at least one unit root	H_0: I(0)
H_1: There is no unit root – i.e. I(0)	H_1: I(1)

- We draw the following conclusions, based on the given combination of results.

ADF Test Result	KPSS Test Result	Conclusion
Reject H_0	Do not reject H_0	The series is I(0)
Do not reject H_0	Reject H_0	The series is I(1)
Reject H_0	Reject H_0	Inconclusive
Do not reject H_0	Do not reject H_0	Inconclusive

Econometric Modeling

i. Non-Stationarity Tests

- Both ln s_t and ln f_t (log-returns) are found to be I(0) – i.e. stationary, as anticipated.

ADF Test for ln s_t

Null Hypothesis: DS has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic based on SIC, MAXLAG=20)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-51.35463	0.0001
Test critical values:	1% level	-3.432171	
	5% level	-2.862230	
	10% level	-2.567181	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(DS)
Method: Least Squares
Date: 04/09/11 Time: 11:16
Sample (adjusted): 3 3256
Included observations: 3254 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
DS(-1)	-0.895622	0.017440	-51.35463	0.0000
C	7.17E-06	0.002728	0.002627	0.9979

R-squared	0.447812	Mean dependent var	7.94E-07
Adjusted R-squared	0.447642	S.D. dependent var	0.209402
S.E. of regression	0.155630	Akaike info criterion	-0.882062
Sum squared resid	78.76522	Schwarz criterion	-0.878321
Log likelihood	1437.116	F-statistic	2637.298
Durbin-Watson stat	1.989725	Prob(F-statistic)	0.000000

KPSS Test for ln s_t

Null Hypothesis: DS is stationary
Exogenous: Constant
Bandwidth: 7 (Newey-West using Bartlett kernel)

		LM-Stat.
Kwiatkowski-Phillips-Schmidt-Shin test statistic		0.080451
Asymptotic critical values*:	1% level	0.739000
	5% level	0.463000
	10% level	0.347000

*Kwiatkowski-Phillips-Schmidt-Shin (1992, Table 1)

Residual variance (no correction)	0.024466
HAC corrected variance (Bartlett kernel)	0.027585

KPSS Test Equation
Dependent Variable: DS
Method: Least Squares
Date: 04/09/11 Time: 11:11
Sample (adjusted): 2 3256
Included observations: 3255 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-1.12E-05	0.002742	-0.004097	0.9967

R-squared	0.000000	Mean dependent var	-1.12E-05
Adjusted R-squared	-0.000000	S.D. dependent var	0.156440
S.E. of regression	0.156440	Akaike info criterion	-0.871981
Sum squared resid	79.63668	Schwarz criterion	-0.870111
Log likelihood	1420.149	Durbin-Watson stat	1.791160

ADF Test for ln f_t

Null Hypothesis: DF has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic based on SIC, MAXLAG=20)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-49.22725	0.0001
Test critical values:	1% level	-3.432171	
	5% level	-2.862230	
	10% level	-2.567181	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(DF)
Method: Least Squares
Date: 04/09/11 Time: 11:18
Sample (adjusted): 3 3256
Included observations: 3254 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
DF(-1)	-0.854011	0.017348	-49.22725	0.0000
C	-0.000438	0.002473	-0.177206	0.8594

R-squared	0.426993	Mean dependent var	-6.86E-06
Adjusted R-squared	0.426817	S.D. dependent var	0.186324
S.E. of regression	0.141064	Akaike info criterion	-1.078591
Sum squared resid	64.71176	Schwarz criterion	-1.074849
Log likelihood	1756.868	F-statistic	2423.323
Durbin-Watson stat	1.997854	Prob(F-statistic)	0.000000

KPSS Test for ln f_t

Null Hypothesis: DF is stationary
Exogenous: Constant
Bandwidth: 5 (Newey-West using Bartlett kernel)

		LM-Stat.
Kwiatkowski-Phillips-Schmidt-Shin test statistic		0.099099
Asymptotic critical values*:	1% level	0.739000
	5% level	0.463000
	10% level	0.347000

*Kwiatkowski-Phillips-Schmidt-Shin (1992, Table 1)

Residual variance (no correction)	0.020315
HAC corrected variance (Bartlett kernel)	0.025660

KPSS Test Equation
Dependent Variable: DF
Method: Least Squares
Date: 04/09/11 Time: 11:19
Sample (adjusted): 2 3256
Included observations: 3255 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.000531	0.002499	-0.212544	0.8317

R-squared	0.000000	Mean dependent var	-0.000531
Adjusted R-squared	-0.000000	S.D. dependent var	0.142552
S.E. of regression	0.142552	Akaike info criterion	-1.057912
Sum squared resid	66.12478	Schwarz criterion	-1.056042
Log likelihood	1722.752	Durbin-Watson stat	1.707886

Econometric Modeling

i. Non-Stationarity Tests

- Both ln S_t and ln F_t are found to be I(1) – i.e. non-stationary, as anticipated.

ADF Test for ln S_t

Null Hypothesis: LOGS has a unit root
Exogenous: Constant
Lag Length: 1 (Automatic based on SIC, MAXLAG=20)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-1.893536	0.3356
Test critical values:	1% level	-3.432171	
	5% level	-2.862230	
	10% level	-2.567181	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(LOGS)
Method: Least Squares
Date: 04/09/11 Time: 11:22
Sample (adjusted): 3 3256
Included observations: 3254 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
LOGS(-1)	-0.001974	0.001042	-1.893536	0.0584
D(LOGS(-1))	0.105365	0.017441	6.041312	0.0000
C	0.009892	0.005224	1.893524	0.0584

R-squared	0.011984	Mean dependent var	7.91E-08
Adjusted R-squared	0.011377	S.D. dependent var	0.001565
S.E. of regression	0.001556	Akaike info criterion	-10.09289
Sum squared resid	0.007868	Schwarz criterion	-10.08728
Log likelihood	16424.13	F-statistic	19.71708
Durbin-Watson stat	1.989852	Prob(F-statistic)	0.000000

KPSS Test for ln S_t

Null Hypothesis: LOGS is stationary
Exogenous: Constant
Bandwidth: 44 (Newey-West using Bartlett kernel)

		LM-Stat.
Kwiatkowski-Phillips-Schmidt-Shin test statistic		1.686200
Asymptotic critical values*:	1% level	0.739000
	5% level	0.463000
	10% level	0.347000

*Kwiatkowski-Phillips-Schmidt-Shin (1992, Table 1)

Residual variance (no correction)	0.000685
HAC corrected variance (Bartlett kernel)	0.029848

KPSS Test Equation
Dependent Variable: LOGS
Method: Least Squares
Date: 04/09/11 Time: 11:24
Sample: 1 3256
Included observations: 3256

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	5.011085	0.000459	10926.76	0.0000

R-squared	-0.000000	Mean dependent var	5.011085
Adjusted R-squared	-0.000000	S.D. dependent var	0.026169
S.E. of regression	0.026169	Akaike info criterion	-4.448197
Sum squared resid	2.229030	Schwarz criterion	-4.446327
Log likelihood	7242.664	Durbin-Watson stat	0.003573

ADF Test for ln F_t

Null Hypothesis: LOGF has a unit root
Exogenous: Constant
Lag Length: 1 (Automatic based on SIC, MAXLAG=20)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-1.798689	0.3816
Test critical values:	1% level	-3.432171	
	5% level	-2.862230	
	10% level	-2.567181	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(LOGF)
Method: Least Squares
Date: 04/09/11 Time: 11:28
Sample (adjusted): 3 3256
Included observations: 3254 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
LOGF(-1)	-0.001654	0.000920	-1.798689	0.0722
D(LOGF(-1))	0.146784	0.017348	8.461165	0.0000
C	0.019732	0.010973	1.798285	0.0722

R-squared	0.022285	Mean dependent var	-5.12E-06
Adjusted R-squared	0.021683	S.D. dependent var	0.001426
S.E. of regression	0.001410	Akaike info criterion	-10.28931
Sum squared resid	0.006465	Schwarz criterion	-10.28370
Log likelihood	16743.71	F-statistic	37.04963
Durbin-Watson stat	1.998113	Prob(F-statistic)	0.000000

KPSS Test for ln F_t

Null Hypothesis: LOGF is stationary
Exogenous: Constant
Bandwidth: 44 (Newey-West using Bartlett kernel)

		LM-Stat.
Kwiatkowski-Phillips-Schmidt-Shin test statistic		1.418622
Asymptotic critical values*:	1% level	0.739000
	5% level	0.463000
	10% level	0.347000

*Kwiatkowski-Phillips-Schmidt-Shin (1992, Table 1)

Residual variance (no correction)	0.000722
HAC corrected variance (Bartlett kernel)	0.031541

KPSS Test Equation
Dependent Variable: LOGF
Method: Least Squares
Date: 04/09/11 Time: 11:29
Sample: 1 3256
Included observations: 3256

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	11.92873	0.000471	25321.55	0.0000

R-squared	0.000000	Mean dependent var	11.92873
Adjusted R-squared	0.000000	S.D. dependent var	0.026881
S.E. of regression	0.026881	Akaike info criterion	-4.394483
Sum squared resid	2.352033	Schwarz criterion	-4.392613
Log likelihood	7155.219	Durbin-Watson stat	0.002811

4) Econometric Modeling

• • •

i. Non-Stationarity Tests
ii. Estimating the Error Correction Model
iii. Estimating the Error Correction Model with Cost of Carry
iv. Estimating the Autoregressive Moving Average Model
v. Estimating the Vector Autoregressive Model
vi. Model Selection

Econometric Modeling

ii. Estimating the Error Correction Model

- According to Equation (1), the spot and futures prices should never drift too far apart, which suggests that the two series might have a cointegrating relationship of the form

$$\ln S_t = \gamma_0 + \gamma_1 \ln F_t \qquad (2)$$

- To test for cointegration, we estimate a regression based on Equation (2) and test the residuals for non-stationarity.

Econometric Modeling

ii. Estimating the Error Correction Model

- The results are inconclusive, as the ADF test finds the residuals to be stationary, whereas the KPSS test does not.

ADF Test for Residuals

Null Hypothesis: RESIDUAL has a unit root
Exogenous: Constant
Lag Length: 12 (Automatic based on SIC, MAXLAG=20)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-6.210651	0.0000
Test critical values:	1% level	-3.432178	
	5% level	-2.862233	
	10% level	-2.567183	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(RESIDUAL)
Method: Least Squares
Date: 04/09/11 Time: 11:57
Sample (adjusted): 14 3256
Included observations: 3243 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
RESIDUAL(-1)	-0.045552	0.007334	-6.210651	0.0000
D(RESIDUAL(-1))	0.147714	0.017875	8.263960	0.0000
D(RESIDUAL(-2))	-0.041834	0.018001	-2.323986	0.0202
D(RESIDUAL(-3))	0.034540	0.017955	1.923683	0.0545
D(RESIDUAL(-4))	-0.022977	0.017966	-1.278884	0.2010
D(RESIDUAL(-5))	-0.025254	0.017929	-1.408545	0.1591
D(RESIDUAL(-6))	-0.444727	0.017885	-24.86547	0.0000
D(RESIDUAL(-7))	-0.031585	0.017511	-1.803777	0.0714
D(RESIDUAL(-8))	-0.021709	0.017475	-1.242291	0.2142
D(RESIDUAL(-9))	0.030591	0.017438	1.754286	0.0795
D(RESIDUAL(-10))	-0.039081	0.017444	-2.240384	0.0251
D(RESIDUAL(-11))	-0.035830	0.017394	-2.059918	0.0395
D(RESIDUAL(-12))	-0.183919	0.017313	-10.62347	0.0000
C	8.24E-06	2.73E-05	0.302145	0.7626

R-squared	0.242564	Mean dependent var	3.81E-06
Adjusted R-squared	0.239515	S.D. dependent var	0.001782
S.E. of regression	0.001554	Akaike info criterion	-10.09198
Sum squared resid	0.007795	Schwarz criterion	-10.06572
Log likelihood	16378.15	F-statistic	79.54374
Durbin-Watson stat	1.990086	Prob(F-statistic)	0.000000

KPSS Test for Residuals

Null Hypothesis: RESIDUAL is stationary
Exogenous: Constant
Bandwidth: 39 (Newey-West using Bartlett kernel)

		LM-Stat.
Kwiatkowski-Phillips-Schmidt-Shin test statistic		6.969195
Asymptotic critical values*:	1% level	0.739000
	5% level	0.463000
	10% level	0.347000

*Kwiatkowski-Phillips-Schmidt-Shin (1992, Table 1)

Residual variance (no correction)	2.16E-05
HAC corrected variance (Bartlett kernel)	0.000469

KPSS Test Equation
Dependent Variable: RESIDUAL
Method: Least Squares
Date: 04/09/11 Time: 11:59
Sample: 1 3256
Included observations: 3256

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-7.70E-16	8.15E-05	-9.45E-12	1.0000

R-squared	0.000000	Mean dependent var	-6.92E-16
Adjusted R-squared	0.000000	S.D. dependent var	0.004653
S.E. of regression	0.004653	Akaike info criterion	-7.902484
Sum squared resid	0.070459	Schwarz criterion	-7.900614
Log likelihood	12866.24	Durbin-Watson stat	0.146144

Econometric Modeling

ii. Estimating the Error Correction Model

- Even though the test for cointegration yielded inconclusive results, we proceed to develop the Error Correction Model (ECM) as if cointegration exists.

- We do this as although the ECM may not be sufficiently robust to be used as the basis of a trading strategy, we develop it as a basis of comparison for the other three models.

** During model selection later, we eventually do not select the ECM. As such, the cointegration assumption here is of no material consequence for the trading strategy.*

Econometric Modeling

ii. Estimating the Error Correction Model

- The ECM can be expressed in the form

$$\Delta \ln S_t = \beta_0 + \delta \hat{z}_{t-1} + \sum_{i=1}^{p} \beta_i \Delta \ln S_{t-i} + \sum_{j=1}^{q} \alpha_j \Delta \ln F_{t-j} + \varepsilon_t \qquad (3)$$

where $\hat{z}_t$ is the t^{th} residual from the regression estimation of Equation (2)

- We develop the ECM by selecting the optimal lags for ln S_t and ln F_t (i.e. p and q), limited to either 1 or 2 lags as according to Abhyankar (1998), the futures price seldom leads the spot price by more than 20 minutes – two 10-minute periods.

Econometric Modeling

ii. Estimating the Error Correction Model

- According to AIC and SBIC, p=1 and q=2.

$$\Delta \ln S_t = \beta_0 + \delta \hat{z}_{t-1} + \beta_1 \Delta \ln S_{t-1} + \alpha_1 \Delta \ln F_{t-1} + \alpha_2 \Delta \ln F_{t-2} + \varepsilon_t \qquad (4)$$

- The AIC and SBIC values for each combination of p and q are below.

		q	
		1	2
p	1	AIC: -10.16769	AIC: -10.17408
		SBIC: -10.16621	SBIC: -10.16473
	2	AIC: -10.16854	AIC: -10.17377
		SBIC: -10.15918	SBIC: -10.16254

Econometric Modeling

ii. Estimating the Error Correction Model

- Then, we fit the ECM based on the first 2,000 observations (the remaining 1,255 are reserved for out-of-sample forecasting later).

Dependent Variable: D1LOGS2000
Method: Least Squares
Date: 04/09/11 Time: 15:30
Sample (adjusted): 3 2000
Included observations: 1998 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.000352	4.20E-05	-8.389011	0.0000
RESIDUAL2000	-0.162238	0.009497	-17.08259	0.0000
D1LOGS2000(-1)	0.141750	0.021629	6.553727	0.0000
D1LOGF2000(-1)	-0.137287	0.025083	-5.473229	0.0000
D1LOGF2000(-2)	-0.148064	0.024533	-6.035265	0.0000

R-squared	0.137703	Mean dependent var	-1.23E-05
Adjusted R-squared	0.135973	S.D. dependent var	0.001776
S.E. of regression	0.001651	Akaike info criterion	-9.972538
Sum squared resid	0.005432	Schwarz criterion	-9.958524
Log likelihood	9967.565	F-statistic	79.56745
Durbin-Watson stat	2.028956	Prob(F-statistic)	0.000000

- We obtain the ECM

$$\Delta \ln S_t = -0.000352 - 0.162\hat{z}_{t-1} + 0.142\Delta \ln S_{t-1} - 0.137\Delta \ln F_{t-1} - 0.148\Delta \ln F_{t-2} \qquad (4)$$

4) Econometric Modeling

• • •

i. Non-Stationarity Tests
ii. Estimating the Error Correction Model
iii. Estimating the Error Correction Model with Cost of Carry
iv. Estimating the Autoregressive Moving Average Model
v. Estimating the Vector Autoregressive Model
vi. Model Selection

Econometric Modeling

iii. Estimating the ECM with Cost of Carry

- The Error Correction Model with cost of carry (ECMCOC) differs from the ECM in that it uses "modified" residuals that incorporate the cost of carry compounded continuously.

$$\hat{z}_t = \ln S_t - \hat{\gamma}_0 - \hat{\gamma}_1 \ln F_t - \hat{\gamma}_2 (r - d)(T - t) \qquad (5)$$

- As with the residuals in the ECM, we test this series for stationarity.

Econometric Modeling

iii. Estimating the ECM with Cost of Carry

- The modified residuals are found to be I(0) – i.e. stationary, as anticipated.

ADF Test for Modified Residuals

Null Hypothesis: RESIDUAL2 has a unit root
Exogenous: Constant
Lag Length: 13 (Automatic based on SIC, MAXLAG=20)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-12.38238	0.0000
Test critical values:	1% level	-3.432179	
	5% level	-2.862233	
	10% level	-2.567183	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(RESIDUAL2)
Method: Least Squares
Date: 04/09/11 Time: 15:37
Sample (adjusted): 15 3256
Included observations: 3242 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
RESIDUAL2(-1)	-0.171062	0.013815	-12.38238	0.0000
D(RESIDUAL2(-1))	0.245104	0.019997	12.25726	0.0000
D(RESIDUAL2(-2))	0.049347	0.019658	2.510230	0.0121
D(RESIDUAL2(-3))	0.121571	0.019457	6.248073	0.0000
D(RESIDUAL2(-4))	0.060418	0.019351	3.122213	0.0018
D(RESIDUAL2(-5))	0.058778	0.019304	3.044803	0.0023
D(RESIDUAL2(-6))	-0.362242	0.019160	-18.90575	0.0000
D(RESIDUAL2(-7))	0.042906	0.020070	2.137801	0.0326
D(RESIDUAL2(-8))	0.025677	0.017750	1.446586	0.1481
D(RESIDUAL2(-9))	0.074059	0.017650	4.196112	0.0000
D(RESIDUAL2(-10))	0.001891	0.017586	0.107556	0.9144
D(RESIDUAL2(-11))	0.005629	0.017565	0.320488	0.7486
D(RESIDUAL2(-12))	-0.149869	0.017437	-8.594663	0.0000
D(RESIDUAL2(-13))	0.058708	0.017616	3.332644	0.0009
C	1.46E-06	2.67E-05	0.054722	0.9564

R-squared	0.267421	Mean dependent var	3.82E-07
Adjusted R-squared	0.264243	S.D. dependent var	0.001773
S.E. of regression	0.001521	Akaike info criterion	-10.13492
Sum squared resid	0.007461	Schwarz criterion	-10.10677
Log likelihood	16443.70	F-statistic	84.14194
Durbin-Watson stat	2.001639	Prob(F-statistic)	0.000000

KPSS Test for Modified Residuals

Null Hypothesis: RESIDUAL2 is stationary
Exogenous: Constant
Bandwidth: 25 (Newey-West using Bartlett kernel)

		LM-Stat.
Kwiatkowski-Phillips-Schmidt-Shin test statistic		0.062937
Asymptotic critical values*:	1% level	0.739000
	5% level	0.463000
	10% level	0.347000

*Kwiatkowski-Phillips-Schmidt-Shin (1992, Table 1)

Residual variance (no correction)	1.14E-05
HAC corrected variance (Bartlett kernel)	6.67E-05

KPSS Test Equation
Dependent Variable: RESIDUAL2
Method: Least Squares
Date: 04/09/11 Time: 15:39
Sample: 1 3256
Included observations: 3256

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.96E-15	5.93E-05	3.30E-11	1.0000

R-squared	-0.000000	Mean dependent var	1.96E-15
Adjusted R-squared	-0.000000	S.D. dependent var	0.003382
S.E. of regression	0.003382	Akaike info criterion	-8.540497
Sum squared resid	0.037227	Schwarz criterion	-8.538628
Log likelihood	13904.93	Durbin-Watson stat	0.273721

Econometric Modeling

iii. Estimating the ECM with Cost of Carry

- We develop the ECMCOC by selecting the optimal lags for ln S_t and ln F_t (i.e. p and q).
- AIC selects p=1 and q=1; while SBIC selects p=2 and q=1. As the differences between the AIC values is very small, we choose p=2 and q=1.

$$\Delta \ln S_t = \beta_0 + \delta \hat{z}_{t-1} + \beta_1 \Delta \ln S_{t-1} + \beta_2 \Delta \ln S_{t-2} + \alpha_1 \Delta \ln F_{t-1} + \varepsilon_t \qquad (6)$$

- The AIC and SBIC values for each pair of p and q are below.

		q	
		1	2
p	1	AIC: -10.25420	AIC: -10.28564
		SBIC: -10.24672	SBIC: -10.27628
	2	AIC: -10.25455	AIC: -10.28555
		SBIC: -10.24519	SBIC: -10.27432

Econometric Modeling

iii. Estimating the ECM with Cost of Carry

- Then, we fit the ECMCOC based on the first 2,000 observations.

Dependent Variable: D1LOGS
Method: Least Squares
Date: 04/09/11 Time: 16:14
Sample (adjusted): 3 2000
Included observations: 1998 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-2.10E-05	3.63E-05	-0.578839	0.5628
RESIDUAL2	-0.202275	0.010505	-19.25484	0.0000
D1LOGS(-1)	0.159009	0.021413	7.425634	0.0000
D1LOGS(-2)	-0.032050	0.020627	-1.553780	0.1204
D1LOGF(-1)	-0.197414	0.025128	-7.856438	0.0000

R-squared	0.168770	Mean dependent var	-1.23E-05
Adjusted R-squared	0.167102	S.D. dependent var	0.001776
S.E. of regression	0.001621	Akaike info criterion	-10.00923
Sum squared resid	0.005236	Schwarz criterion	-9.995212
Log likelihood	10004.22	F-statistic	101.1631
Durbin-Watson stat	2.062694	Prob(F-statistic)	0.000000

- We obtain the ECM

$$\Delta \ln S_t = -0.22\hat{z}_{t-1} + 0.142\Delta \ln S_{t-1} - 0.137\Delta \ln F_{t-1} - 0.148\Delta \ln F_{t-2} \qquad (7)$$

4) Econometric Modeling

• • •

i. Non-Stationarity Tests
ii. Estimating the Error Correction Model
iii. Estimating the Error Correction Model with Cost of Carry
iv. Estimating the Autoregressive Moving Average Model
v. Estimating the Vector Autoregressive Model
vi. Model Selection

Econometric Modeling

iv. Estimating the Autoregressive Moving Average Model

- The ARMA estimates spot prices from historical prices with white noise. It takes the form of

$$y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \theta_1 u_{t-1} + \theta_2 u_{t-2} + \ldots + \theta_q u_{t-q} + u_t$$

where y_t is ln S_t

u_t is the t^{th} error term

- We develop the ARMA by selecting the optimal lags for ln S_t and u_t (i.e. p and q).

Econometric Modeling

iv. Estimating the Autoregressive Moving Average Model

- Based on SBIC, we choose p=1 and q=1.

$$\ln S_t = \mu + \Phi_1 \ln S_{t-1} + \theta_1 u_{t-1} + u_t$$

- The SBIC values for each pair of p and q are below.

		q		
		0	1	2
p	0	-	0.148913	0.141839
	1	4.214903	0.140214	0.143038
	2	3.199872	0.142408	0.146642

Econometric Modeling

iv. Estimating the Autoregressive Moving Average Model

- Then, we fit the ARMA based on the first 2,000 observations.

$$\ln S_t = -0.2012 + 0.9136 \ln S_{t-1} + 0.1245 u_{t-1} + u_t$$

4) Econometric Modeling

• • •

i. Non-Stationarity Tests
ii. Estimating the Error Correction Model
iii. Estimating the Error Correction Model with Cost of Carry
iv. Estimating the Autoregressive Moving Average Model
v. Estimating the Vector Autoregressive Model
vi. Model Selection

Econometric Modeling

v. Estimating the Vector Autoregressive Model

- A VAR differs from the other models in that it is a systems regression model – i.e. there is more than one dependent variable.
- We develop a simple bivariate VAR of the form

$$s_t = \beta_{10} + \beta_{11}\, s_{t-1} + \ldots + \beta_{1k}\, s_{t-k} + \alpha_{11}\, f_{t-1} + \ldots \alpha_{1k}\, f_{t-k} + u_{1t}$$

$$f_t = \beta_{20} + \beta_{21}\, s_{t-1} + \ldots + \beta_{2k}\, s_{t-k} + \alpha_{21}\, f_{t-1} + \ldots \alpha_{2k}\, f_{t-k} + u_{2t}$$

- We develop the VAR by selecting the optimal number of lags.

Econometric Modeling

v. Estimating the Vector Autoregressive Model

- AIC selects 14 lags, HQIC selects 13 and SBIC selects 7.

Lag	LogL	LR	FPE	AIC	SC	HQ
0	12389.39	NA	1.26e-08	-12.51251	-12.50687	-12.51044
1	19834.17	14867.00	6.87e-12	-20.02845	-20.01151	-20.02223
2	19911.43	154.1358	6.38e-12	-20.10245	-20.07422	-20.09208
3	19936.64	50.23347	6.24e-12	-20.12387	-20.08434	-20.10935
4	19962.68	51.85259	6.11e-12	-20.14614	-20.09532	-20.12747
5	20013.92	101.9192	5.82e-12	-20.19386	-20.13174	-20.17104
6	20019.89	11.85837	5.81e-12	-20.19585	-20.12244	-20.16888
7	20344.48	644.2613	4.20e-12	-20.51968	-20.43497	-20.48856
8	20348.16	7.285898	4.20e-12	-20.51935	-20.42335	-20.48408
9	20349.21	2.094868	4.22e-12	-20.51638	-20.40908	-20.47696
10	20353.74	8.953937	4.21e-12	-20.51691	-20.39831	-20.47334
11	20359.61	11.60382	4.21e-12	-20.51880	-20.38891	-20.47108
12	20364.53	9.710765	4.20e-12	-20.51972	-20.37854	-20.46786
13	20395.38	60.87177	4.09e-12	-20.54685	-20.39437	-20.49084
14	20399.87	8.846341	4.09e-12*	-20.54735	-20.38357	-20.48718
15	20402.64	5.440329	4.09e-12	-20.54610	-20.37103	-20.48178
16	20404.82	4.288803	4.10e-12	-20.54426	-20.35790	-20.47580
17	20410.09	10.36921	4.09e-12	-20.54555	-20.34789	-20.47294
18	20410.43	0.649066	4.11e-12	-20.54184	-20.33289	-20.46508
19	20415.66	10.26363	4.10e-12	-20.54309	-20.32285	-20.46218
20	20422.22	12.85717*	4.09e-12	-20.54568	-20.31414	-20.46062

Econometric Modeling

v. Estimating the Vector Autoregressive Model

- However, as explained in the paper, a modified multivariate criteria from Enders (1995) was used rather than simple multivariate criteria, such that we proceed to build the VAR with 1 lag.

- We obtain the VAR

$$\ln s_t = -0.857191 + 0.851429 \ln s_{t-1} + 0.134239 \ln f_{t-1} + u_{1t}$$

$$\ln f_t = -0.128885 + 1.021868 \ln f_{t-1} + -0.026358 \ln s_{t-1} + u_{2t}$$

Econometric Modeling

v. Estimating the Vector Autoregressive Model

- Granger causality implies correlation between the current value of a variable and the past values of other variables
- F-test jointly tests for the significance of the lags on the explanatory variables

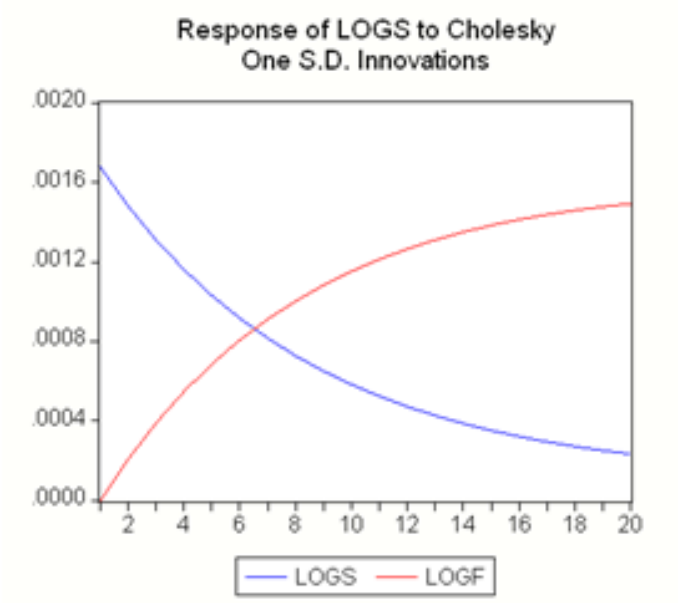
Dependent Variable: LOGF			
Excluded	**Chi-Square**	**df**	**Probability**
LOGS	8.485380	1	0.0036
All	8.485380	1	0.0036

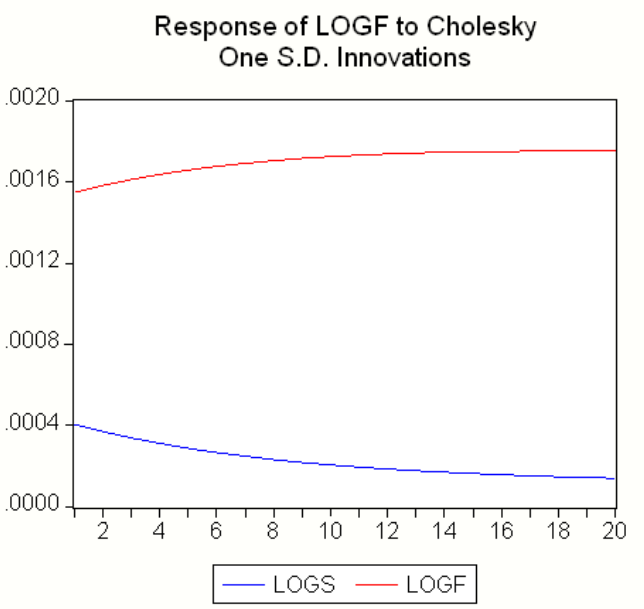
Dependent Variable: LOGS			
Excluded	**Chi-Square**	**df**	**Probability**
LOGF	243.8957	1	0.0000
All	243.8957	1	0.0000

Econometric Modeling

v. Estimating the Vector Autoregressive Model

- The impulse response functions can be used to produce the time path of the dependent variables in the VAR, to shocks from all the explanatory variables.

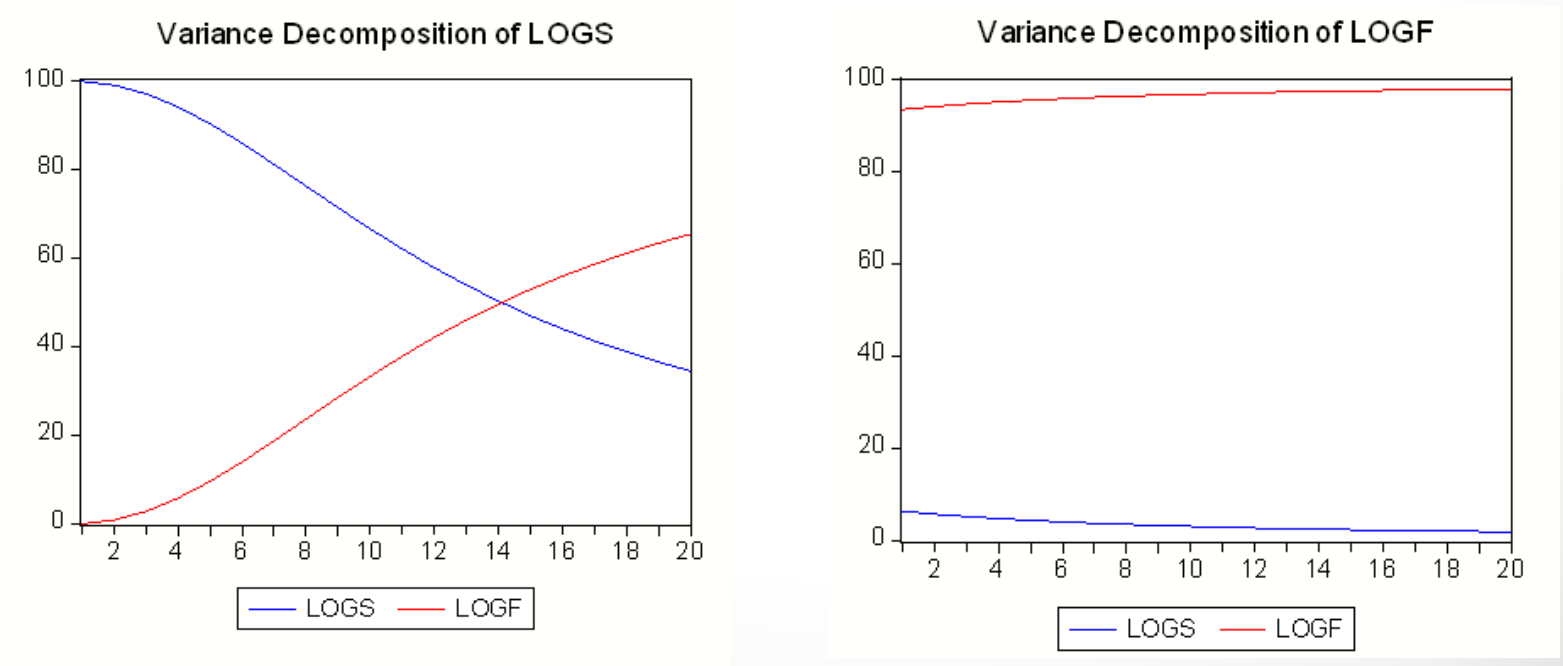




Econometric Modeling

v. Estimating the Vector Autoregressive Model

- Variance decomposition also examines the effects of shocks to dependent variables, by determining how much of the forecast error variance is explained by innovations to each independent variable, over a series of time horizons.



4) Econometric Modeling

• • •

i. Non-Stationarity Tests
ii. Estimating the Error Correction Model
iii. Estimating the Error Correction Model with Cost of Carry
iv. Estimating the Autoregressive Moving Average Model
v. Estimating the Vector Autoregressive Model
vi. Model Selection

Econometric Modeling

vi. Model Selection

- Each of the four models was fitted based on the first 2,000 observations.

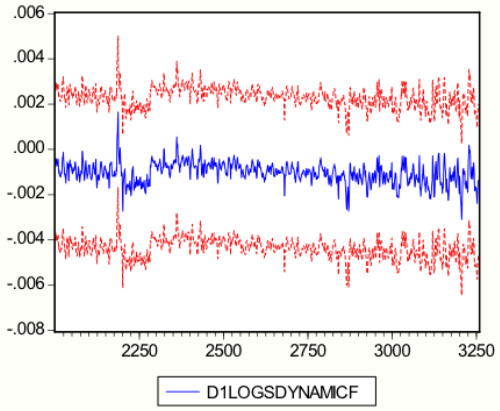
- To select the model to be used as the basis for the trading strategies later, we use the fitted models to forecast the next 1,256 values and then compare them with the 1,256 remaining observations.

Econometric Modeling

vi. Model Selection

- The forecasts are as follows

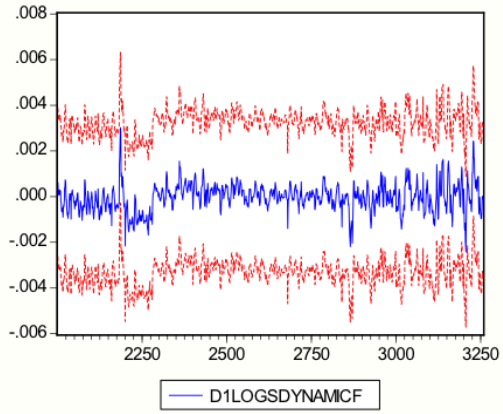
ECM



Forecast: D1LOGSDYNAMICF
Actual: D1LOGS
Forecast sample: 2001 3256
Included observations: 1255

Root Mean Squared Error	0.001498
Mean Absolute Error	0.001182
Mean Abs. Percent Error	657.1650
Theil Inequality Coefficient	0.652997
Bias Proportion	0.501042
Variance Proportion	0.205973
Covariance Proportion	0.292984

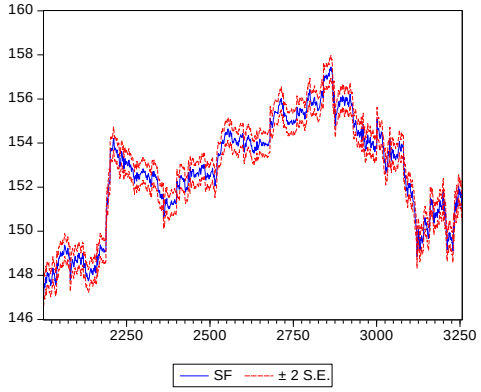
ECMCOC



Forecast: D1LOGSDYNAMICF
Actual: D1LOGS
Forecast sample: 2001 3256
Included observations: 1255

Root Mean Squared Error	0.001091
Mean Absolute Error	0.000726
Mean Abs. Percent Error	292.2664
Theil Inequality Coefficient	0.623889
Bias Proportion	0.003011
Variance Proportion	0.257801
Covariance Proportion	0.739188

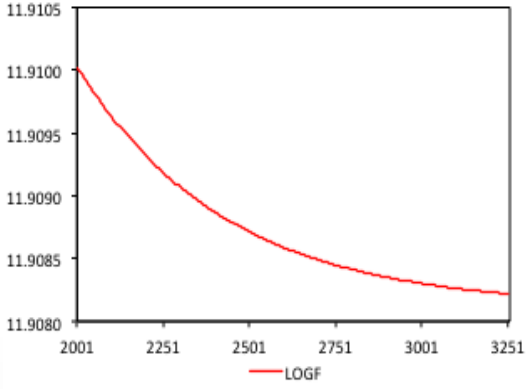
ARMA



Forecast: SF
Actual: S
Forecast sample: 2001 3256
Included observations: 1256

Root Mean Squared Error	0.174480
Mean Absolute Error	0.115177
Mean Abs. Percent Error	0.075576
Theil Inequality Coefficient	0.000571
Bias Proportion	0.007667
Variance Proportion	0.000075
Covariance Proportion	0.992258

VAR



Forecast: LOGF

Forecast sample: 2001 1256
Included observations: 1256

Root Mean Squared Error	0.038095
Mean Absolute Error	0.03432

Econometric Modeling

vi. Model Selection

- Based on the forecasting errors of the models, we select the ECMCOC as it has the smallest errors.

Model	Root Mean Squared Error	Mean Absolute Error
ECM	0.001498	0.001182
ECMCOC	0.001091	0.000726
ARMA	0.174480	0.115177
VAR	0.038095	0.034320

5) Formulating a Trading Strategy

• • •

i. Description of 8 Trading Strategies
ii. Trading Simulation Environment and Assumptions
iii. Comparison of Simulation Results

5) Formulating a Trading Strategy

• • •

i. Description of 8 Trading Strategies
ii. Trading Simulation Environment and Assumptions
iii. Comparison of Simulation Results

Formulating a Trading Strategy

i. Description of 8 Trading Strategies

- Strategy 1: Liquidity trading strategy
  - Trading on the basis of every positive predicted return and making a round trip trade. If return is predicted to be negative, no trade will be made.
- Strategy 2: Buy and hold strategy
  - Trading based on every positive predicted return and hold the position until the next return is predicted to be negative. This strategy attempts to reduce the amount of transaction costs.
- Strategy 3: Filter strategy – better than predicted average
  - Trading only if predicted returns is larger than average predicted return, which is calculated to be 0.000659676, and hold the position unit the next return is predicted to be negative. Similarly, this strategy attempts to reduce the amount of transaction costs.

Formulating a Trading Strategy

i. Description of 8 Trading Strategies

- Strategy 4: Filter strategy – better than predicted first decile
  - Trading only if predicted returns is larger than the first decile predicted return, which is calculated to be 0.001434563, and hold the position unit the next return is predicted to be negative.
- Strategy 5: Filter strategy – high arbitrary cutoff
  - Trading only if predicted returns is larger than a high arbitrary cut-off point, which is 0.0022, and hold the position unit the next return is predicted to be negative.
- Strategy 6: Passive investment
  - Buy at the start of the out-sample trading period and sell only at the end of the out-sample trading period.

Formulating a Trading Strategy

i. Description of 8 Trading Strategies

- Strategy 7: Filter strategy – search for 1-tier dynamic filter
  - Dynamically search for 1 cutoff point that yields the best returns from the in-sample data, which is calculated to be 0.001005. Trading only if the predicted return is larger than this cutoff point, and hold the position unit the next return is predicted to be negative.
- Strategy 8: Filter strategy – search for 2-tier dynamic filter
  - Dynamically search for 2 cutoff points that yields the best returns from the in-sample data, which is calculated to be 0.001 and 0.001001. Trade 1 lot if the predicted return is larger than the first cutoff point, and trade another lot if the predicted return is larger than the second cutoff point. Sell off one lot if the predicted return falls below the second cutoff point, and sell off all holdings if the next return is predicted to be negative.

5) Formulating a Trading Strategy

• • •

i. Description of 8 Trading Strategies
ii. Trading Simulation Environment and Assumptions
iii. Comparison of Simulation Results

Formulating a Trading Strategy

ii. Trading Simulation Environment and Assumptions

- Initial portfolio value is $1000
- Transaction cost, which includes commission, stamp duty and bid-ask spread is assumed to be 0.3% of the ETF price for each buy or sell transaction
- Each strategy trades and holds a maximum of 2 lots of ETF at any point in time

5) Formulating a Trading Strategy

• • •

i. Description of 8 Trading Strategies
ii. Trading Simulation Environment and Assumptions
iii. Comparison of Simulation Results

Formulating a Trading Strategy

iii. Comparison of Simulation Results

- As expected, Liquidity Trading strategy trades the most number of transactions
- Buy and Hold is the best strategy when transaction costs are ignored
- Better than predicted first decile filter strategy is the best strategy when transaction costs are considered.

Strategy	Number of Transactions	Portfolio Value without Transaction Costs	Portfolio Value with Transaction Costs
Liquidity trading	2548	1065.19	-102.12
Buy and hold	344	1065.19	907.55
Filter average	100	1046.02	1000.37
Filter decile	12	1013.84	1008.42
Filter high cutoff	8	1010.91	1007.29
Passive investment	4	1007.17	1005.38
1-tier dynamic filter	40	1023.41	1005.25
2-tier dynamic filter	40	1024.04	1005.87

6) Conclusion

• • •

i. Areas for Improvement
ii. Overall Conclusions

6) Conclusion

• • •

i. Areas for Improvement
ii. Overall Conclusions

Conclusion

i. Areas for Improvement

1. One area of improvement is to use tick-by-tick bid and ask quotes instead of tick-by-tick transaction data. We noticed that there may not be any transactions for both ETF and Futures during every 10 minute period. Hence, using bid and ask quotes will ensure that the data is continuous. Also, using bid and ask quotes will factor in the exact bid and ask spread as transaction cost.

2. Another area of improvement is to use more recent data for simulation. There are many data vendors who can provide more recent data for a fee.

Conclusion

i. Areas for Improvement

3. The reason for choosing S&P 500 index for our experiment is because S&P 500 is one of the more popular index in the financial markets. Another area of improvement is to try out other popular indices such as Dow Jones Industrial Average, to find out which index could be more profitable.

4. The reason for choosing SPDR S&P 500 ETF (SPY) is because it is the first and most popular ETF in USA. However, this ETF will still have some tracking error. Another area of improvement is to search for a better S&P 500 ETF with a low tracking error to replace SPY, which will improve our simulation results.

Conclusion

i. Areas for Improvement

5. The ECMCOC is the best model in terms of predictive ability. However, the optimized coefficients are always changing as confirmed by checking using out-sample data. Hence, another area of improvement is to dynamically check the optimized coefficients and adjust the trading strategies for changes.

Dependent Variable: D1LOGS
Method: Least Squares
Date: 04/14/11 Time: 20:10
Sample: 2001 3255
Included observations: 1255

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2.77E-05	3.02E-05	0.916586	0.3595
RESIDUAL2	-0.157397	0.011579	-13.59289	0.0000
D1LOGS(-1)	0.122603	0.027604	4.441445	0.0000
D1LOGF(-1)	-0.126989	0.031233	-4.065829	0.0001

R-squared	0.138105	Mean dependent var	1.92E-05
Adjusted R-squared	0.136039	S.D. dependent var	0.001151
S.E. of regression	0.001070	Akaike info criterion	-10.83859
Sum squared resid	0.001433	Schwarz criterion	-10.82222
Log likelihood	6805.216	F-statistic	66.81791
Durbin-Watson stat	2.024953	Prob(F-statistic)	0.000000

6) Conclusion

• • •

i. Areas for Improvement

ii. Overall Conclusions

Conclusion

ii. Overall Conclusions

- Our experiment investigated the lead-lag relationship between the S&P 500 index and futures prices and confirmed that the futures returns lead the spot returns.
- The best model in terms of predictive ability is the Error Correction Model with cost of carry (ECM-COC).
- In the absence of transaction costs, the "Buy and Hold" strategy derived from the ECM-COC model is the most profitable strategy.
- Considering transaction costs, the "Better than predicted first decile filter" strategy is the most profitable strategy.

Conclusion

ii. Overall Conclusions

- In our experiment, we attempted to dynamically search for the best 1-tier filter cut-off point and the best 2-tier filter cut-off points using the in-sample data, and then simulate the 2 trading strategies using the out-sample data. Both strategies yield positive profits, but they are still lower than the profit generated from the passive investment strategy.

Conclusion

ii. Overall Conclusions

- The "lead-lag" relationship between the Spot and Futures is likely due to the following reasons:
  - Some components of the index are infrequently traded, implying that the observed index value contains "stale" component prices.
  - It is more expansive to transact in the spot market (in our experiment, we are using an ETF to represent the spot market) and hence, the spot market reacts more slowly to news.
  - Stock market indices are recalculated only every minute so that new information takes longer to be reflected in the index.

Conclusion

ii. Overall Conclusions

- Our simulation results suggest that we may earn higher profits over the passive investment strategy as shown by the "Better than predicted first decile filter" strategy . However, we are not able to replicate such profits using dynamically searching methods. Hence, this suggests that we may not always profit from the "lead-lag" relationship between the Spot and Futures, and their existence is largely consistent with the absence of arbitrage opportunities and is in accordance with modern definitions of the efficient markets hypothesis.

End

• • •

Thank You