ETFs AND THEIR IMPACT ON BID-ASK SPREAD

By

Achyut Gautam

Abstract

The study explores the relationship between ETF ownership and the bid-ask spread of the underlying securities. The paper tests the hypothesis that the greater the volume of ETF trade, the wider the bid-ask spread of its underlying. In particular, the study tests the impact of SPDR and IVV trading on the S&P 500 and QQQ and TQQQ trading on NASDAQ-100. Using econometric analysis, the paper corroborates its primary findings and concludes that there is a strong significant positive relationship between the bid-ask spread of S&P 500 and its ETFs while the same holds true but only for NASDAQ-100 and QQQ.

I. Introduction

Active investors allocate a lot of resources to scrutinize the financial market and encourage higher standards of corporate governance and efficient capital allocation. However, for several years, active fund managers have failed to "beat the market" and there is an ongoing debate over the advantages and disadvantages of actively managed portfolio. There are several studies that explore cost and benefit to investors who choose to invest in passive or active funds.

The growth in passive fund investment has been steadily increasing over the last two decades. According to the Bank for International Settlement (BIS), within the US, passive funds' estimated share of total outstanding securities is about 15%. Although passive funds presence in the financial market remains relatively low compared to actively managed funds, BIS mention that they grew by 138% from 2007-2017. Exchange Traded Funds, commonly known as ETFs, is a type of passive fund that tracks a stock index, commodities, bonds, or a basket of assets. According to a research conducted by Ernst and Young (EY), at the end of September 2017, global ETF assets totaled to $4.4 trillion - a cumulative average growth rate (CAGR) of about 21% from 2015. The U.S. is the world's largest ETF market and as of 2017, the U.S. ETF industry alone had roughly $3 trillion in total industry asset (Muphy). With trillions of dollars of assets invested in ETFs, ETFs undoubtedly play a significant role in financial markets and their influence is increasing at an impressive rate. Thus, it is important to understand their impact on the aggregate stock market as stock markets play a crucial role in shaping economies.

The goal of this paper is to investigate whether an increase in the volume of ETF trade leads to an increase in bid-ask spread, a transaction cost, of the underlying. The paper addresses this issue by analyzing the liquidity of ETFs and their impact on the bid-ask spread of the underlying. The paper examines SPDR and IVV, ETFs that track S&P 500, and QQQ and TQQ, ETFs that track NASDAQ – 100. Since ETFs are easier to trade and are cost-effective, ETFs encourages passive investment leading to lesser scrutiny from the market. Based on this rationale, following hypothesis is developed:

H_0: An increase in ETF ownership increases bid-ask spread of the underlying
H_a: An increase in ETF ownership does not increase the bid-ask spread of the underlying

II. Literature Review

There are a number of studies that look at the effects of ETFs on different stock market components. Glosten, S.Nallareddy, and Y. Zou (2015) study the effect of ETF trading on information efficiency. They find that ETF trading increases the information efficiency for small firms and firms with imperfect competitive capital markets by timely incorporating accounting information into stock prices. They find that systematic fundamental information rather than firm-specific fundamental information increases informational efficiency. Ben-David, Franzoni, and Moussawi (2017) investigate whether ETFs increase trade volatility and conclude that the stocks that are traded as ETFs reflect higher volatility than otherwise similar securities. Consequently, noise in stock prices increases with ETF ownership. Their findings also show that ETFs themselves create a noise in the market as opposed to simply just acting as a vessel for noise traders. Israeli, Lee, and Sridharan (2017) look at the relationship between ETFs trade, returns, and pricing efficiency. They find that while ETF trading might improve pricing discovery for same-quarter macro-based earnings, in the long run, it degrades the earnings. Further, their findings suggest that an increase in ETFs trades leads to an increase in trading cost which is captured by an

increase in bid-ask spread.

This paper takes a similar approach to that of Israeli et.al in that it uses the high-low spread as a proxy for the bid-ask spread. The study differs from that of Israeli et.al because the underlying is the entire stock market index – S&P 500 and Nasdaq - 100 – as opposed to particular stocks. This study also uses two different measure of bid-ask spread to explore the relationship. Lastly, the study focuses only on econometric methods to provide empirical evidence.

III. Data

In order to analyze this relationship, bi-variate and multivariate regression models are used. The data used in this paper is drawn from FACTSET and Yahoo Finance. Below are the three models that are used to analyze the transaction cost of both the S&P 500 and NASDAQ – 100:

Model 1 - Simple Linear Regression

$$HLSPREAD\_SP500 = \beta_0 + \beta_1 \%\Delta ETF_{SPDR,vol} + \varepsilon \quad (1.1)$$
$$HLSPREA\_NASDAQ = \beta_0 + \beta_1 \%\Delta ETF_{QQQ,vol} + \varepsilon \quad (1.2)$$

Model 2 - Multiple Linear Regression

$$HLSPREAD\_SP500 = \beta_0 + \beta_1 \%\Delta ETF_{SPDR,vol} + \beta_2 \%\Delta ETF_{IVV,vol} + \beta_3 Volatility_{VIX} + \varepsilon \quad (2.1)$$
$$HLSPREA\_NASDAQ = \beta_0 + \beta_1 \%\Delta ETF_{QQQ,vol} + \beta_2 \%\Delta ETF_{TQQQ,vol} + \beta_3 Volatility_{VXN} + \varepsilon \quad (2.2)$$

Model 3 - Multiple Linear Regression Using Corwin and Schultz High-Low

$$HLSPREAD\_SP500_{CORWIN} = \beta_0 + \beta_1 \%\Delta ETF_{SPDR,vol} + \beta_2 \%\Delta ETF_{IVV,vol} + \beta_3 Volatility_{VIX} + \varepsilon \quad (3.1)$$
$$HLSPREAD\_NASDAQ_{CORWIN} = \beta_0 + \beta_1 \%\Delta ETF_{TQQQ,vol} + \beta_2 \%\Delta ETF_{QQQ,vol} + \beta_3 Volatility_{VXN} + \varepsilon \quad (3.2)$$

Simple Linear Regression

1. *High-Low Spread*

   The dependent variable *High-Low Spread (HLSPREAD)* is the monthly high-low measure of bid-ask spread for a stock market index over a month. *HLSPREAD* in Model 1 and Model 2 is obtained in the following way:

   $$HLSPREAD_{i,t} = \frac{HIGH\ (i,t) - LOW\ (i,t)}{LOW\ (i,t)}$$

   where *i* represents the stock market index and *t* represents a particular time period.

HLSPREAD_ i $_{\text{Corwin}}$ used in Model 3 is the Corwin and Schultz (2012) annual high-low a measure of bid-ask spread. Corwin and Schultz (2012) derive the formula in the following way:

$$S = \frac{2(e^{\alpha} - 1)}{1 + e^{\alpha}} \tag{4}$$

$$\alpha = \frac{\sqrt{2\beta} - \beta}{3 - 2\sqrt{2}} - \sqrt{\frac{\delta}{3 - 2\sqrt{2}}} \tag{5}$$

$$\beta = \sum_{j=0}^{i} \left[\ln\left(\frac{H^{\circ}_{t+j}}{L^{\circ}_{t+j}}\right)\right]^2 \tag{6}$$

$$\delta = \left[\ln\left(\frac{H^{\circ}_{t,t+1}}{L^{\circ}_{t,t+1}}\right)\right]^2 \tag{7}$$

The Corwin-Schultz bid-ask spread estimator is represented by equation (4) above. (5) represents the difference between the adjustment of a single month and a 2-month period. (6) represents the monthly high and low price adjustments to the high price, and (7) represents 2-month period high and low adjustments. As mentioned by Israeli et. al, Corwin-Schultz measure of bid-ask spread is less time and data-intensive and they have demonstrated that their measure outperforms the Roll (1984), Lesmond et al. (1999), and Holden (2009) techniques for measuring bid-ask spreads.

2. *ETF* $_{m,\ vol}$

   Δ *% ETF* $_{m,\ vol}$ is the monthly change in the percentage of *m* ETF held by all investors. The Δ operator indicates a change in the value of a respective variable. Based on the hypothesis, the coefficient β_1 is expected to be positive suggesting that an increase in ETF trade increases bid-ask spread of the underlying.

Multiple Linear Regression

3. *ETF* $_{n,\ vol}$

   Δ *% ETF* $_{n,\ vol}$ is the monthly change in the percentage of *n* ETF held by all investors. β_2 is expected to be positive suggesting that an increase in ETF trade increases bid-ask spread of the underlying.

4. *Volatility*

   The CBOE Volatility Index (VIX) and the CBOE Nasdaq Volatility Index (VXN) are used as proxies for the *Volatility*. VIX measures the market's expectations of volatility implied by the S&P 500 over the coming 30 days and VXN measures the market's expectations of 30-day volatility for NASDAQ-100. β_3, the coefficient of *Volatility*, is expected to be positive as during a rapid market change bid-ask spread is much wider as market makers have greater opportunities to take advantage of it.

Classical Linear Assumptions

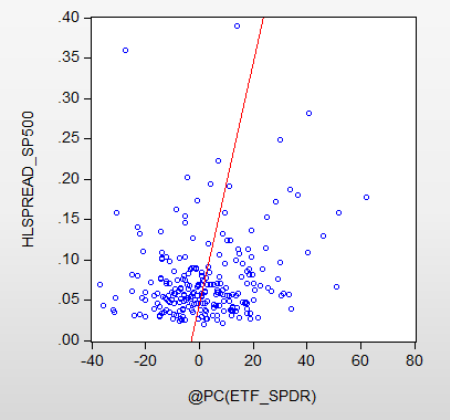
The first assumption states that the parameters must be linear themselves, correctly specified and has an additive error term. This assumption is validated in the results section. The second assumption states that the error term has a zero population mean. Data used in this study are empirical results obtained from actual trades that occurred in random order and the samples were randomly taken from the population for the regression analysis. Similarly, all the models include y-intercept, β_0, which observes the non-zero mean of the error term. These two reasons fulfills the second assumption. Gauss Markov's third assumption tells that regressors being calculated should not be perfectly correlated with each other. As evident in the multicollinearity test, no two explanatory variables are perfectly collinear. Thus, there is no problem of multicollinearity. The assumption of zero conditional mean states that there are no omitted variables that have an effect on the explanatory variables. This assumption is violated if relevant variables are left out or if irrelevant variables are incorporated. Apart from the control variables, results from omitted variable test shows that are no omitted variables that significantly impact dependent variables. The fourth assumption is the assumption of exogeneity which tells that regressors should not be correlated with the error term. Results of serial correlation suggest that there is no serial correlation. Due to the lack of data for control variables such as institutional ownership and number of analysts analyzing the underlying, the study was unable to prove normality assumption of the error terms and despite using autoregressive conditional heteroscedasticity and autoregressive distributed lag model the study was unable reject the presence of heteroscedasticity.

IV. Result

A. Model 1 - Simple Linear Regression - Eq (1.1)

$$HLSPREAD\_SP500 = \beta_0 + \beta_1 \%\Delta \text{ETF}_{\text{SPDR}} + \varepsilon \qquad (1.1)$$

Scatter Plot



Regression Result

Dependent Variable: HLSPREAD_SP500
Method: Least Squares
Date: 03/23/19 Time: 15:12
Sample (adjusted): 1999M02 2018M12
Included observations: 239 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.072026	0.003238	22.24580	0.0000
@PC(ETF_SPDR)	0.000667	0.000200	3.325567	0.0010

R-squared	0.044584	Mean dependent var	0.073309
Adjusted R-squared	0.040552	S.D. dependent var	0.050737
S.E. of regression	0.049698	Akaike info criterion	-3.157384
Sum squared resid	0.585357	Schwarz criterion	-3.128292
Log likelihood	379.3074	Hannan-Quinn criter.	-3.145661
F-statistic	11.05940	Durbin-Watson stat	0.760739
Prob(F-statistic)	0.001022		

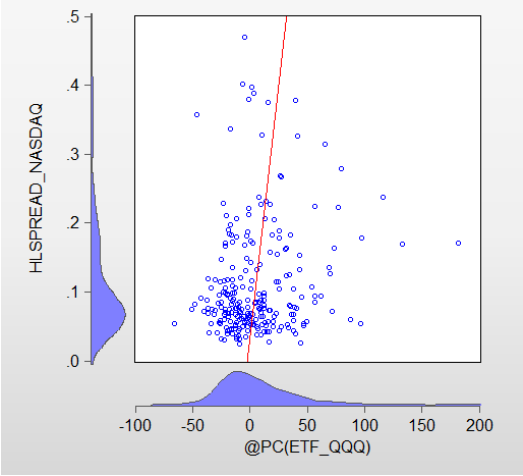
Table 1: Regression for eq (1.1)

The coefficient of *SPDR* is statistically significant and its sign matches the theoretical expectation. The coefficient of 0.000667 tells that with every % increase in the volume of SPDR, the bid-ask spread of S&P 500 increases by 0.067%. R^2 = 0.044584 suggests that the model explains about 4.4584% of the variance around the dependent variable. Prob. of F stats = 0.0% shows that the slope of explanatory variable ≠ 0 and the value of R^2 is statistically significant.

B. Model 1 - Simple Linear Regression - Eq (1.2)

$$HLSPREAD\_NASDAQ = \beta_0 + \beta_1 \%\Delta ETF_{QQQ,\,vol} + \varepsilon \qquad (1.2)$$

Scatter Plot



Regression Result

Dependent Variable: HLSPREAD_NASDAQ
Method: Least Squares
Date: 03/23/19 Time: 15:11
Sample (adjusted): 1999M04 2018M12
Included observations: 237 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.108785	0.005314	20.47206	0.0000
@PC(ETF_QQQ)	0.000440	0.000163	2.701903	0.0074

R-squared	0.030129	Mean dependent var	0.111137
Adjusted R-squared	0.026002	S.D. dependent var	0.081771
S.E. of regression	0.080700	Akaike info criterion	-2.187742
Sum squared resid	1.530453	Schwarz criterion	-2.158475
Log likelihood	261.2474	Hannan-Quinn criter.	-2.175945
F-statistic	7.300279	Durbin-Watson stat	0.468769
Prob(F-statistic)	0.007397		

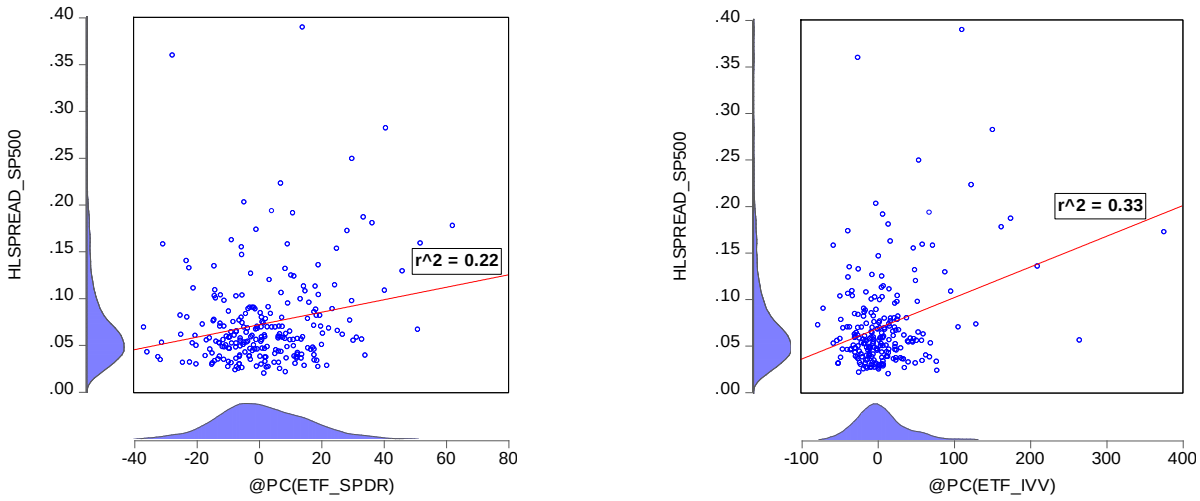
Table 2: Regression for eq (1.2)

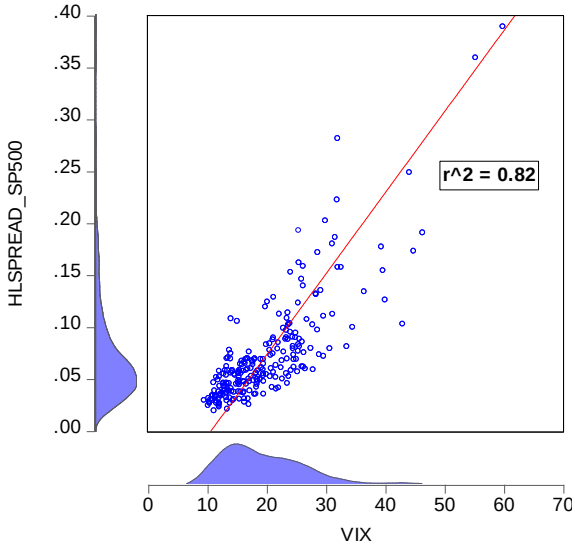
The coefficient of *QQQ* is statistically significant and its sign matches the theoretical expectation. The coefficient of 0.00044 tells that with every % increase in the volume of QQQ, the bid-ask spread of NASDAQ increases by 0.044%. R^2 = 0.030129 suggests that the model explains about 3.0129% of the variance around the dependent variable. Prob. of F stats = 0.0% tells that the slope of explanatory variable $\neq 0$ and the value of R^2 is statistically significant.

C. Multiple Linear Regressions – Model 2 – Eq (2.1)

$$HLSPREAD\_SP500 = \beta_0 + \beta_1 \%\Delta ETF_{SPDR,\,vol} + \beta_2 \%\Delta ETF_{IVV,\,vol} + \beta_3 Volatility_{VIX} \quad (2.1)$$

Scatter Plot Diagram





Summary Statistics

	HLSPREAD_SP500	@PC(ETF_SPDR)	@PC(ETF_IVV)	VIX
Mean	0.072232	1.917630	9.994218	19.60834
Median	0.057387	0.561516	0.623211	17.40000
Maximum	0.389652	62.29720	376.2476	59.89000
Minimum	0.020003	-36.73825	-77.65860	9.510000
Std. Dev.	0.051891	16.31646	50.77734	8.069725
Skewness	2.814040	0.559846	3.022201	1.782316
Kurtosis	13.94807	3.890541	18.18497	7.444795
Jarque-Bera	1408.017	19.01793	2481.973	301.6336
Probability	0.000000	0.000074	0.000000	0.000000
Sum	16.10776	427.6315	2228.711	4372.660
Sum Sq. Dev.	0.597768	59102.33	572391.0	14456.74
Observations	223	223	223	223

Regression Output

Dependent Variable: HLSPREAD_SP500
Method: Least Squares
Date: 03/23/19 Time: 13:17
Sample (adjusted): 2000M06 2018M12
Included observations: 223 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.030777	0.004799	-6.412817	0.0000
@PC(ETF_SPDR)	0.000250	0.000126	1.977495	0.0492
@PC(ETF_IVV)	0.000146	4.11E-05	3.544833	0.0005
VIX	0.005155	0.000229	22.49610	0.0000

R-squared	0.730813	Mean dependent var	0.072232
Adjusted R-squared	0.727126	S.D. dependent var	0.051891
S.E. of regression	0.027106	Akaike info criterion	-4.360323
Sum squared resid	0.160911	Schwarz criterion	-4.299208
Log likelihood	490.1760	Hannan-Quinn criter.	-4.335651
F-statistic	198.1874	Durbin-Watson stat	2.144074
Prob(F-statistic)	0.000000		

Table 3: Regression for eq (1.2)

The coefficient of *SPDR, IVV,* and *VIX* are all statistically significant and their sign matches the theoretical expectation. The model explains 72.7% of the variation around the dependent variable. Adjusted R^2 has increased to 72.7% compared to a model with eq. (1.1) suggesting that the additional explanatory variables helped better explain the variation around the dependent variable. Prob. of F stats = 0.0% shows that the slope of explanatory variable $\neq 0$ and the value of R^2 is statistically significant.

Multicollinearity

The correlation between independent variables < 0.5 and centered VIF < 5. This shows that there is no strong presence of multicollinearity.

Correlation

	VIX	@PC(ETF_SPDR)	@PC(ETF_IVV)
VIX	1.000000	0.082580	0.178818
@PC(ETF_SPDR)	0.082580	1.000000	0.468885
@PC(ETF_IVV)	0.178818	0.468885	1.000000

Variance Inflation Factor (VIF)

Variable	Coefficient Variance	Uncentered VIF	Centered VIF
C	2.30E-05	6.990495	NA
@PC(ETF_SPDR)	1.59E-08	1.299598	1.281813
@PC(ETF_IVV)	1.69E-09	1.366302	1.315124
VIX	5.25E-08	7.159794	1.033034

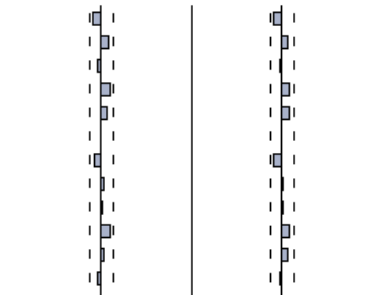
Serial Correlation

The autocorrelations and partial autocorrelations all lags are zero and *Q*-statistics are insignificant with *p*-values > 10%. Breusch – Godfrey Serial Correlation LM Test is insignificant which also suggest that there is no serial correlation. Lastly, DW stats lies in the no serial correlation region. All three results suggest that there is no serial correlation in the residuals.

Correlogram and Q statistics

Correlogram of Residuals

Date: 03/23/19 Time: 11:03
Sample: 1999M01 2018M12
Included observations: 223



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.095	-0.095	2.0218	0.155
		2	0.084	0.076	3.6420	0.162
		3	-0.034	-0.020	3.9110	0.271
		4	0.101	0.091	6.2290	0.183
		5	0.072	0.095	7.4220	0.191
		6	-0.003	-0.003	7.4235	0.283
		7	-0.080	-0.091	8.9213	0.258
		8	0.035	0.015	9.2020	0.326
		9	0.018	0.020	9.2792	0.412
		10	0.100	0.093	11.620	0.311
		11	0.032	0.068	11.856	0.375
		12	-0.025	-0.022	12.003	0.445

Breusch – Godfrey Serial Correlation LM Test

Breusch-Godfrey Serial Correlation LM Test:
Null hypothesis: No serial correlation at up to 2 lags

F-statistic	1.699230	Prob. F(2,217)	0.1852
Obs*R-squared	3.438576	Prob. Chi-Square(2)	0.1792

Durbin Watson Statistics

DW Stats $_{S\&P500}$ = 2.144074

Omitted Variable Test

The probability that omitted variables are not significant is less than 0.0% for *vix and IVV and 4% for SPDR*. This suggests that all the explanatory variables are relevant in the model.

VIX

Omitted Variable Test
Null hypothesis: VIX is not significant
Equation: UNTITLED
Specification: HLSPREAD_SP500 C @PC(ETF_SPDR) @PC(ETF_IVV)
Omitted Variables: VIX

	Value	df	Probability
t-statistic	22.49610	219	0.0000
F-statistic	506.0747	(1, 219)	0.0000
Likelihood ratio	266.9763	1	0.0000

IVV

Omitted Variable Test
Null hypothesis: @PC(ETF_IVV) is not significant
Equation: UNTITLED
Specification: HLSPREAD_SP500 C @PC(ETF_SPDR) VIX
Omitted Variables: @PC(ETF_IVV)

	Value	df	Probability
t-statistic	3.544833	219	0.0005
F-statistic	12.56584	(1, 219)	0.0005
Likelihood ratio	12.44173	1	0.0004

SPDR

Omitted Variable Test
Null hypothesis: @PC(ETF_SPDR) is not significant
Equation: UNTITLED
Specification: HLSPREAD_SP500 C @PC(ETF_IVV) VIX
Omitted Variables: @PC(ETF_SPDR)

	Value	df	Probability
t-statistic	1.977495	219	0.0492
F-statistic	3.910485	(1, 219)	0.0492
Likelihood ratio	3.946776	1	0.0470

Unit Root Test

Conducting unit root test in first difference (as well as in level and second difference) shows that the probability of having a unit root is less than 0.001% for both dependent and independent variables.

HLSPREAD_SP500

Augmented Dickey-Fuller Unit Root Test on D(HLSPREAD_SP500)

Null Hypothesis: D(HLSPREAD_SP500) has a unit root
Exogenous: Constant
Lag Length: 2 (Automatic - based on SIC, maxlag=14)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-13.34050	0.0000
Test critical values:	1% level	-3.457984	
	5% level	-2.873596	
	10% level	-2.573270	

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

SPDR

Augmented Dickey-Fuller Unit Root Test on D(ETF_SPDR)

Null Hypothesis: D(ETF_SPDR) has a unit root
Exogenous: Constant
Lag Length: 3 (Automatic - based on SIC, maxlag=14)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-12.08098	0.0000
Test critical values:	1% level	-3.458104	
	5% level	-2.873648	
	10% level	-2.573298	

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

IVV

Augmented Dickey-Fuller Unit Root Test on D(ETF_IVV)

Null Hypothesis: D(ETF_IVV) has a unit root
Exogenous: Constant
Lag Length: 1 (Automatic - based on SIC, maxlag=14)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-15.16480	0.0000
Test critical values:	1% level	-3.459898	
	5% level	-2.874435	
	10% level	-2.573719	

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

VIX

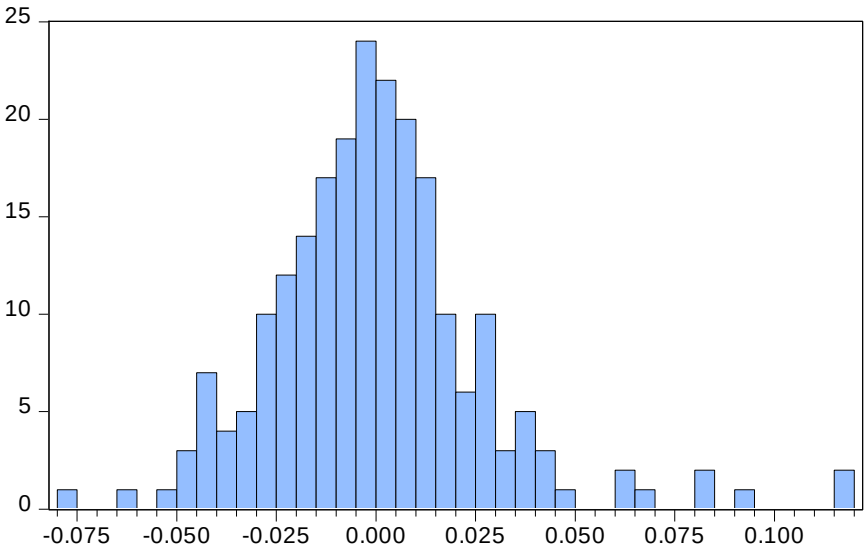
Augmented Dickey-Fuller Unit Root Test on D(VIX)

Null Hypothesis: D(VIX) has a unit root
Exogenous: Constant
Lag Length: 1 (Automatic - based on SIC, maxlag=14)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-13.08905	0.0000
Test critical values:	1% level	-3.457865	
	5% level	-2.873543	
	10% level	-2.573242	

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

Normality Test



Series: Residuals
Sample 2000M06 2018M12
Observations 223

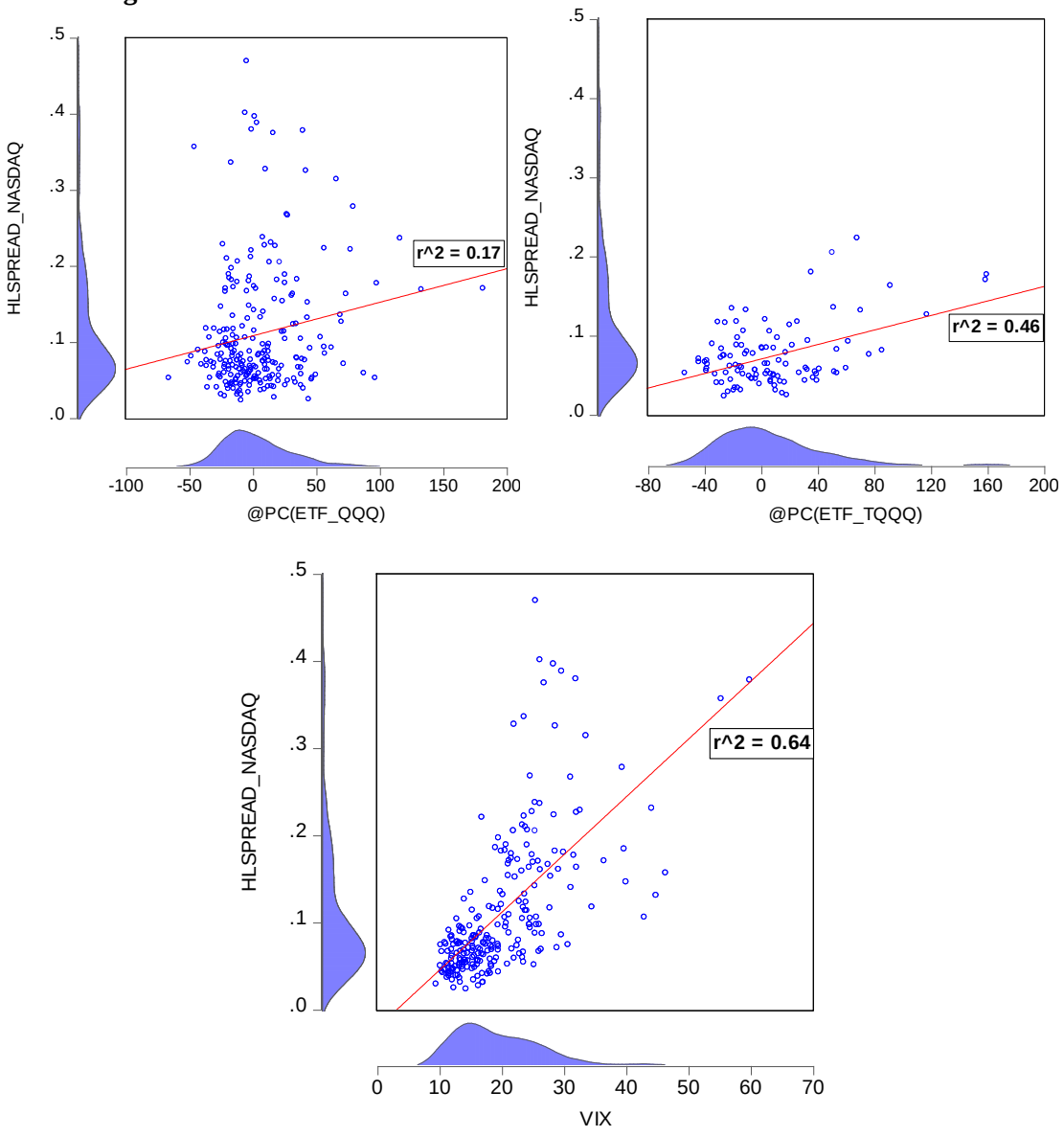
Mean	-2.36e-17
Median	-0.000856
Maximum	0.116066
Minimum	-0.079013
Std. Dev.	0.026923
Skewness	0.976295
Kurtosis	6.308700
Jarque-Bera	137.1460
Probability	0.000000

The normality test for the residual fails in this model despite having large n and shows that the errors do not follow a normal distribution. Although OLS does not require error terms to follow normal distribution to produce unbiased estimates with minimum variance, satisfying normality test generates more reliable confidence intervals.

D. Multiple Linear Regressions – Model 2 – Eq (2.2)

$$HLSPREAD\_NASDAQ = \beta_0 + \beta_1 \%\Delta ETF_{QQQ,vol} + \beta_2 \%\Delta ETF_{TQQQ,vol} + \beta_3 Volatility_{VXN} + \varepsilon \quad (2.2)$$

Scatter Plot Diagram



Summary Statistics

	HLSPREAD_NASDAQ	@PC(ETF_QQQ)	@PC(ETF_TQQQ)	VXN
Mean	0.074580	4.797253	7.757704	19.21764
Median	0.062673	-4.491070	1.940934	17.64500
Maximum	0.223770	181.4832	159.6822	44.98000
Minimum	0.024248	-66.14962	-53.85048	11.53000
Std. Dev.	0.038697	35.80225	38.90812	5.497016
Skewness	1.578766	1.681960	1.492310	1.659652
Kurtosis	5.663835	7.817459	6.124308	6.895813
Jarque-Bera	75.37497	152.4804	82.45589	115.6952
Probability	0.000000	0.000000	0.000000	0.000000
Sum	7.905439	508.5088	822.3166	2037.070
Sum Sq. Dev.	0.157230	134589.1	158953.4	3172.805
Observations	106	106	106	106

Regression Result

Dependent Variable: HLSPREAD_NASDAQ
Method: Least Squares
Date: 03/24/19 Time: 14:28
Sample (adjusted): 2010M03 2018M12
Included observations: 106 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.005529	0.010489	-0.527094	0.5993
@PC(ETF_QQQ)	4.33E-05	0.000130	0.333241	0.7396
@PC(ETF_TQQQ)	0.000219	0.000126	1.739785	0.0849
VXN	0.004069	0.000539	7.551012	0.0000

R-squared	0.498681	Mean dependent var	0.074580
Adjusted R-squared	0.483936	S.D. dependent var	0.038697
S.E. of regression	0.027799	Akaike info criterion	-4.290647
Sum squared resid	0.078823	Schwarz criterion	-4.190140
Log likelihood	231.4043	Hannan-Quinn criter.	-4.249911
F-statistic	33.82104	Durbin-Watson stat	2.145212
Prob(F-statistic)	0.000000		

Table 4: Regression for eq (2.2)

The coefficient of *TQQQ a*nd *VIX* are statistically significant and their sign matches the theoretical expectation. However, although *QQQ* has the correct sign it is not statistically significant. This issue will be dealt with later under the omitted variable test. The model explains 72.7% of the variation around the dependent variable. Adjusted R^2 has increased to 72.7% compared to the model with eq. (1.1) suggesting that the additional explanatory variables helped explain the variation around the dependent variable better. Prob. of F stats = 0.0% shows that the slope of explanatory variable $\neq 0$ and the value of R^2 is statistically significant.

Multicollinearity Test

The correlation between independent variables < 0.5 and centered VIF < 5. This shows that there is no strong presence of multicollinearity.

Variance Inflation Factors
Date: 03/23/19 Time: 14:28
Sample: 1999M01 2018M12
Included observations: 106

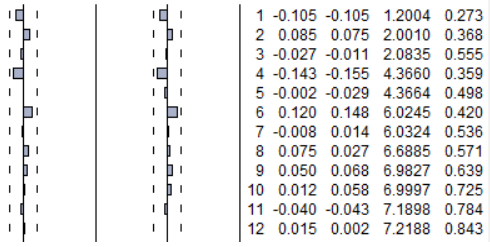
Variable	Coefficient Variance	Uncentered VIF	Centered VIF
C	0.000110	15.09120	NA
@PC(ETF_QQQ)	1.68E-08	2.987139	2.933961
@PC(ETF_TQQQ)	1.58E-08	3.388187	3.257456
VXN	2.90E-07	15.90485	1.192401

Serial Correlation Test

The autocorrelations and partial autocorrelations all lags are zero and *Q*-statistics are insignificant with *p*-values > 10%. Breusch – Godfrey Serial Correlation LM Test is insignificant which also suggest that there is no serial correlation. Lastly, DW stats lies in the no serial correlation region. All three results suggest that there is no serial correlation in the residuals.

Correlogram and Q statistics

Correlogram of Residuals
Date: 03/23/19 Time: 11:06
Sample: 1999M01 2018M12
Included observations: 106



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.105	-0.105	1.2004	0.273
		2	0.085	0.075	2.0010	0.368
		3	-0.027	-0.011	2.0835	0.555
		4	-0.143	-0.155	4.3660	0.359
		5	-0.002	-0.029	4.3664	0.498
		6	0.120	0.148	6.0245	0.420
		7	-0.008	0.014	6.0324	0.536
		8	0.075	0.027	6.6885	0.571
		9	0.050	0.068	6.9827	0.639
		10	0.012	0.058	6.9997	0.725
		11	-0.040	-0.043	7.1898	0.784
		12	0.015	0.002	7.2188	0.843

Breusch – Godfrey Serial Correlation LM Test

Breusch-Godfrey Serial Correlation LM Test:
Null hypothesis: No serial correlation at up to 2 lags

F-statistic	0.973280	Prob. F(2,100)	0.3814
Obs*R-squared	2.023956	Prob. Chi-Square(2)	0.3635

Durbin Watson Statistics

DW Stats $_{NASDAQ}$ = 2.145212

Omitted Variable Test

The probability that *VXN* is not significant is less than 0.0%. However, the probability that *TQQQ* and *QQQ* is not significant is more than 5%.

VXN

Omitted Variable Test
Null hypothesis: VXN is not significant
Equation: UNTITLED
Specification: HLSPREAD_NASDAQ C @PC(ETF_QQQ) @PC(ETF_TQQQ)
Omitted Variables: VXN

	Value	df	Probability
t-statistic	7.551012	102	0.0000
F-statistic	57.01778	(1, 102)	0.0000
Likelihood ratio	47.06858	1	0.0000

QQQ

Omitted Variable Test
Null hypothesis: @PC(ETF_QQQ) is not significant
Equation: UNTITLED
Specification: HLSPREAD_NASDAQ C @PC(ETF_TQQQ) VXN
Omitted Variables: @PC(ETF_QQQ)

	Value	df	Probability
t-statistic	0.333241	102	0.7396
F-statistic	0.111049	(1, 102)	0.7396
Likelihood ratio	0.115341	1	0.7341

TQQQ

Omitted Variable Test
Null hypothesis: @PC(ETF_TQQQ) is not significant
Equation: UNTITLED
Specification: HLSPREAD_NASDAQ C @PC(ETF_QQQ) VXN
Omitted Variables: @PC(ETF_TQQQ)

	Value	df	Probability
t-statistic	1.739785	102	0.0849
F-statistic	3.026853	(1, 102)	0.0849
Likelihood ratio	3.099784	1	0.0783

Since *QQQ* and *TQQQ* both together are not relevant in the model, the following model with eq. (2.3) excludes *TQQQ*. *TQQQ* is removed because it has less observation than *QQQ* and removing *TQQQ* increases adjusted R^2, fixes the omitted variable bias, and the significance level of *QQQ*.

Omitted Variable Test
Null hypothesis: @PC(ETF_QQQ) is not significant
Equation: UNTITLED
Specification: HLSPREAD_NASDAQ C VXN
Omitted Variables: @PC(ETF_QQQ)

	Value	df	Probability
t-statistic	3.062925	216	0.0025
F-statistic	9.381510	(1, 216)	0.0025
Likelihood ratio	9.311039	1	0.0023

$$HLSPREAD\_SP500 = \beta_0 + \beta_1 \%\Delta \text{ETF}_{QQQ,\,vol} + \beta_3 \text{Volatility}_{VXN} \quad (2.3)$$

Dependent Variable: HLSPREAD_NASDAQ
Method: Least Squares
Date: 03/23/19 Time: 18:31
Sample (adjusted): 2000M10 2018M12
Included observations: 219 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.021188	0.005099	-4.155184	0.0000
@PC(ETF_QQQ)	0.000237	7.75E-05	3.062925	0.0025
VXN	0.004739	0.000172	27.58364	0.0000

R-squared	0.786956	Mean dependent var	0.103955
Adjusted R-squared	0.784984	S.D. dependent var	0.078207
S.E. of regression	0.036264	Akaike info criterion	-3.782359
Sum squared resid	0.284062	Schwarz criterion	-3.735934
Log likelihood	417.1683	Hannan-Quinn criter.	-3.763609
F-statistic	398.9383	Durbin-Watson stat	1.823475
Prob(F-statistic)	0.000000		

Table 5: Regression for eq (2.3)

Unit Root Test

Conducting unit root test in first difference (as well as in level and second difference) shows that the probability of having a unit root is less than 0.001% for both dependent and independent variables.

HLSPREAD_NASDAQ

Augmented Dickey-Fuller Unit Root Test on D(HLSPREAD_NASDAQ)

Null Hypothesis: D(HLSPREAD_NASDAQ) has a unit root
Exogenous: Constant
Lag Length: 3 (Automatic - based on SIC, maxlag=14)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-12.28377	0.0000
Test critical values:	1% level	-3.458104	
	5% level	-2.873648	
	10% level	-2.573298	

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

QQQ

Augmented Dickey-Fuller Unit Root Test on D(ETF_QQQ)

Null Hypothesis: D(ETF_QQQ) has a unit root
Exogenous: Constant
Lag Length: 3 (Automatic - based on SIC, maxlag=14)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-11.77594	0.0000
Test critical values:	1% level	-3.458347	
	5% level	-2.873755	
	10% level	-2.573355	

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

TQQQ

Augmented Dickey-Fuller Unit Root Test on D(ETF_TQQQ)

Null Hypothesis: D(ETF_TQQQ) has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic - based on SIC, maxlag=12)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-11.36623	0.0000
Test critical values:	1% level	-3.493747	
	5% level	-2.889200	
	10% level	-2.581596	

VXN

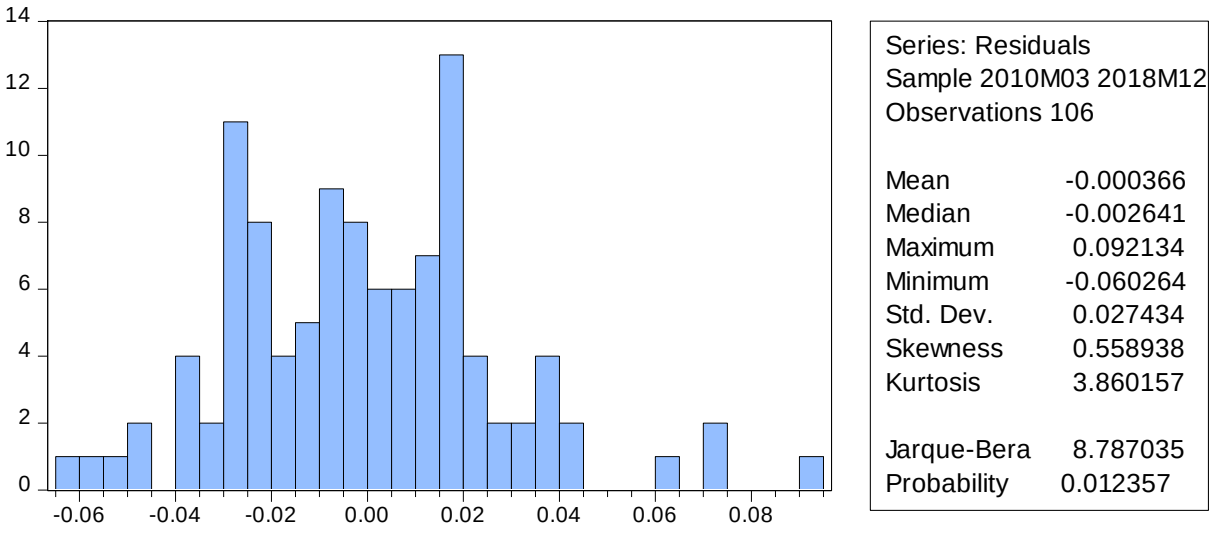
Augmented Dickey-Fuller Unit Root Test on D(VXN)

Null Hypothesis: D(VXN) has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic - based on SIC, maxlag=14)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-16.03815	0.0000
Test critical values:	1% level	-3.460453	
	5% level	-2.874679	
	10% level	-2.573850	

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

Normality Test



The normality test for the residual fails in this model despite having large n. This shows that the errors do not follow a normal distribution and violates the seventh assumption of classical Linear regression.

E. Model 3 - Multiple Linear Regression Using Corwin and Schultz High-Low Spread

Using Corwin and Schultz bid-ask spread estimator (2012) as a dependent variable led to independent variables *SPDR, IVV, and TQQQ* to have a wrong coefficient sign. Unlike the model used by Israeli et.al, models below do not account for control variables. This might have led to the wrong direction of the variables.

$$HLSPREAD\_SP500_{CORWIN} = \beta_0 + \beta_1 \%\Delta ETF_{SPDR,vol} + \beta_2 \%\Delta ETF_{IVV,vol} + \beta_3 Volatility_{VIX} + \varepsilon \quad (3.1)$$

Dependent Variable: HLSPREAD_SP500_CORWIN
Method: Least Squares
Date: 03/23/19 Time: 16:27
Sample (adjusted): 2000M06 2018M12
Included observations: 223 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.000460	0.003694	-0.124584	0.9010
@PC(ETF_SPDR)	-0.000164	9.72E-05	-1.690546	0.0923
@PC(ETF_IVV)	-6.43E-05	3.16E-05	-2.034005	0.0432
VIX	0.000908	0.000176	5.151265	0.0000

Table 6: Regression for eq (3.1)

$$HLSPREAD\_NASDAQ_{CORWIN} = \beta_0 + \beta_1 \%\Delta ETF_{QQQ,vol} + \beta_2 \%\Delta ETF_{TQQQ,vol} + \beta_3 Volatility_{VXN} + \varepsilon \quad (3.2)$$

Dependent Variable: HLSPREAD_NASDAQ_CORWIN
Method: Least Squares
Date: 03/24/19 Time: 13:08
Sample (adjusted): 2010M03 2018M12
Included observations: 106 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.020434	0.006560	-3.114803	0.0024
@PC(ETF_QQQ)	7.66E-05	8.12E-05	0.943911	0.3474
@PC(ETF_TQQQ)	-0.000218	7.87E-05	-2.765313	0.0068
VXN	0.002061	0.000337	6.115402	0.0000

R-squared	0.277078	Mean dependent var	0.017857
Adjusted R-squared	0.255816	S.D. dependent var	0.020154
S.E. of regression	0.017386	Akaike info criterion	-5.229251
Sum squared resid	0.030833	Schwarz criterion	-5.128744
Log likelihood	281.1503	Hannan-Quinn criter.	-5.188515
F-statistic	13.03137	Durbin-Watson stat	1.751386
Prob(F-statistic)	0.000000		

Table 6: Regression for eq (3.2)

V. Conclusion

As hypothesized in this paper and from the results obtained, it can be observed that an increase in ETF trade leads to the wider bid-ask spread of the underlying. More specifically, the multivariate regressions with eq. 2.1 and 2.3 explain more than 70% of the variation of the bid-ask spread of its underlying at 5% significance level. The paper excludes control variables such as institutional ownership and number of analysts analyzing the underlying due to lack of data. Incorporating control variables and effectively controlling for noise might have increased the robustness and efficiency of the models and established homoscedasticity and normality assumption.

Works Cited

Ben-David, Itzhak, Francesco Franzoni, and Rabih Moussawi. "Do ETFs increase volatility?." *The Journal of Finance* 73.6 (2018): 2471-2535.

Corwin, Shane A., and Paul Schultz. "A simple way to estimate bid-ask spreads from daily high and low prices." *The Journal of Finance* 67.2 (2012): 719-760.

Grant, Turner, and Vladyslav Sushko. "The implication of passive investing for securities markets." *The Journal of Finance* 73.6 (2018): 2471-2535.

Sushko, Vladyslav, and Grant Turner. "The Implications of Passive Investing for Securities Markets." *The Bank for International Settlements*, 11 Mar. 2018, www.bis.org/publ/qtrpdf/r_qt1803j.htm. [Accessed 15 Mar. 2019]

Glosten, Lawrence, Suresh Nallareddy, and Yuan Zou. *ETF trading and informational efficiency of underlying securities*. Working paper. Columbia University. October, 2015.

Gujarati, Damodar N., and Dawn C. Porter. *Basic Econometrics*. McGraw-Hill/Irwin, 2017.

Israeli, Doron, Charles MC Lee, and Suhas A. Sridharan. "Is there a dark side to exchange traded funds? An information perspective." *Review of Accounting Studies* 22.3 (2017): 1048-1083.

Julie, Kerr, Lisa Kealy, and Matt Forstenhausler. *Global ETF Research.* Ernst & Young Global Limited. Available at: www.ey.com/industries/financial-services/asset-management/ey-global-etf-survey-2017 [Accessed 15 Mar. 2019].

Murphy, Cinthia. "What US ETF Market Looks Like Today." What US ETF Market Looks Like Today | ETF.com, 13 Apr. 2017, www.etf.com/sections/features-and-news/what-us-etf-market-looks-today