Understanding GameStop volatility through Reddit comments

Akbar Qaiser Amin
University of California, Irvine

ECON 137W, Summer 2021

Abstract

This study examines the short squeeze on GameStop that exploited the short positions of institutional investors. It is likely that the events were initiated by retail investors on the subreddit, r/wallstreetbets. This paper investigates the impact of change in the WSB subreddit's attention towards the short squeeze on the volatility of GameStop's stock. A WSB interest indicator is created to measure the subreddit's comment activity for a particular term relative to all of Reddit. Contrary to what is expected, the strength of relationship between term interest and the next day's price variation is weak.

Keywords: GameStop, Reddit, wallstreetbets, short squeeze.

1. Introduction

Early 2021, small-scale investors concentrated around the subreddit r/wallstreetbets (WSB) bet against institutional investors and coordinated a short squeeze[1] on several underrated stocks, including GameStop. As a brief overview, r/wallstreetbets is a subreddit—or discussion forum—on the Reddit website where participants discuss stock and option trading. While the subreddit is well-known for its reckless trading strategies, its use of strong and colorful language stands out the most. Phrases are coined for almost anything related to finance—investors holding short positions are known as *gay bears*, those who are risk-averse have *paper hands*, while those who adamantly hold a position have *diamond hands*. The profane and juvenile nature of the subreddit help describe the general intention of the movement: to oppose those who control a significant portion of a market and make profit—or in WSB lingo, *go to the moon*. The David-vs-Goliath narrative was fitting for the movement and is what led those with no financial literacy to suddenly want to buy underrated stocks like GameStop. The reason to why GameStop had such a large short float[2] in the first place is not the main topic of this paper; rather, it is to explore the relationship between the colorful vocabulary of the WSB subreddit and the price variation, or volatility, of GameStop.

In this study, I provide evidence on how the activity on the WSB subreddit for some terms drove daily price variations for GameStop's stock. Specifically, I adapt the Google search term indicator (Rui X., 2015) and instead measure the level of interest in GameStop on the WSB subreddit and the level of interest for all of Reddit through comments. Regressions are run for

---

[1] A short-squeeze occurs when short-sellers rapidly move to cover their positions for a stock, buying more than the relative market volume. Intuitively, the stock price will rise, but if the stock is heavily shorted, other short-sellers are pressured to do the same, resulting in a cascade of stock purchases.

[2] The average number of days short sellers take to cover a short position, i.e., a large a short float implies short sellers hold their short position for a longer period compared to those with smaller short floats.

various volatility windows with the expectation that a shorter duration should be easier to predict. If price variations are driven by the WSB subreddit, an increase in the ratio should result in a higher price variation for GameStop. I expect this to hold for terms relating specifically to GameStop, such as its ticker symbol, but perhaps not for other terms.

2. Data & Regression

In this project, I used two main datasets-

- Daily quote data for GameStop (NYSE:GME). The data was obtained using Yahoo! Finance and contains the open, close, high, and low prices for each day.
- The total number of comments on the WSB subreddit and all of Reddit containing terms related to GameStop or meme stocks by day. The comments are queried programmatically via HTTP requests to Reddit's Application Programming Interface (API) using Python.

All data are sampled daily within the period starting from January 2021 to April 2021. Although the exact date the activity on the WSB subreddit started is difficult to determine, the chosen period should be long enough to capture the initial activities.

The data from mentioned sources had to be pre-processed in order to be suitable for reliable analysis. While Reddit comment data was available for all days, prices for GameStop were understandably absent for weekends and holidays. In order to complete this data, the missing the values were imputed using linear interpolation. Given the value of GameStop on a given date is x and the next available data point is y, the missing values in-between are approximated recursively using the concave function $\frac{x+y}{2}$. Daily price variation for GameStop is estimated using the Yang-Zhang volatility indicator (Yang and Zhang, 2000). Since the estimator requires a sample period N, multiple regressions are run with varying sample size. Let o_t, h_t, l_t, and c_t be the open, high, low, and close prices at time t. The Yang-Zhang volatility for an asset i is:

$$V_t^i = \sqrt{\sigma_o^2 + k\sigma_c^2 + (1-k)\sigma_{rs}^2}$$

where:

$$\sigma_c^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(ln\frac{c_i}{o_{i-1}}\right)^2,$$

$$\sigma_o^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(ln\frac{o_i}{c_{i-1}}\right)^2,$$

$$k = 0.34/1.34 + \frac{N+1}{N-1},$$

and σ_{rs}^2 is the Rogers and Satchell (1991) estimator defined as:

$$\sigma_{rs}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(ln\frac{h_i}{c_i}\right)\left(ln\frac{h_i}{o_i}\right) + \left(ln\frac{l_i}{c_i}\right)\left(ln\frac{l_i}{o_i}\right)$$

The Rogers-Satchell estimator does not consider price movements between trading sessions; however, the Yang-Zhang estimator is more comprehensive in that it considers open-session jumps.

The terms for which comments are counted include: *gme*, *yolo*[3], *tendies*[4], and *stonks*[5]. The last three terms do not share any significance with; however, they do describe the spirit of the subreddit's movement, i.e., to profit from risky trading strategies. Single words are

[3] Acronym for "you only live once". It is used to rationalize reckless behavior.
[4] Refers for chicken tenders (loosely speaking); however, is used to describe financial gain.
[5] Deliberate misspelling of stocks. Used to humorously to imply poor investment decisions.

intentionally chosen to avoid dealing with the complex nature of multi-word phrases. For example, the term *short squeeze is* frequently used on the WSB subreddit in various contexts, such as "I hear there might be another squeeze" and "Not every stock is a short squeeze." While all queries are interpreted by the Reddit API as case-insensitive, there is no way to query an exact case and/or exclude specific cases (in contrast to SQL queries). Therefore, it did not seem appropriate to generalize the significance of a term based on a single use case.

Let WSB_t^i and $Reddit_t^i$ denote the number of occurrences of term i in comments for day t for the WSB subreddit and all of Reddit, respectively. The WSB subreddit interest indicator is:

$$W_t^i = ln\left(\frac{WSB_t^i}{Reddit_t^i}\right)$$

Figure 1 visualizes the untransformed ratio between WSB and Reddit term occurrences. The natural logarithm is used to account for the nonlinear nature of the events, i.e., the data needs to be normalized. This follows the practices of a similar study (Rui X, 2015) done on Google search trends.

The goal of this paper is to evaluate the impact of change in WSB activity towards specific terms on the next day price variation for GameStop. To study this, I use a linear regression model estimated via OLS of the following form:

$$V_t^{\mathrm{GME}} = \beta_0 + \beta_1\left(V_{t-1}^{\mathrm{GME}}\right) + \beta_3\left(W_{t-1}^{GME}\right) + \beta_4\left(W_{t-1}^{YOLO}\right) + \beta_5\left(W_{t-1}^{STONKS}\right) + \beta_6\left(W_{t-1}^{TENDIES}\right) + \epsilon_t$$

where β are the explanatory variables, V_t^{GME} and $\mathrm{V}_{\mathrm{t-1}}^{\mathrm{GME}}$ are the current (t) and lagged $(t-1)$ price variations for GameStop, W_{t-1}^i is the WSB subreddit interest for a selected term i and ϵ_{t} is the error term accounting for the variation in GameStop that cannot be explained by the variation in explanatory variables.

3. Results

Tables 1 – 4 show the corresponding regression results for the 3-day, 7-day, 14-day, and 21-day volatility windows, respectively. Equivalently, Figure 2 shows a Q-Q plot of the residuals against quantiles of the t-distribution with 6 degrees of freedom. From the Q-Q- plot alone, the data points curving off from the 45-degree line imply that the term occurrences data contained many extreme values, i.e. they have a “heavy tail.”

For all windows, the 1-day variation lag for GameStop (V_{t-1}^{GME}) strongly correlates to the price variation in GameStop (V_t^{GME}) and is statistically significant at 95% (two-tailed test), where the critical values range from 14 to 65. Consistently across results, the WSB indicator for the term *tendies* ($W_{t-1}^{TENDIES}$) has a negative correlation. This agrees with my expectation that the WSB indicator may not be sufficient for non-GameStop related terms. In fact, the term for the ticker symbol (W_{t-1}^{GME}) along with *yolo* (W_{t-1}^{YOLO}) hold slightly positive coefficients for the first two windows. Similarly, as the window increases, the relationship between the terms and the next day’s price variation weakens, which follows my expectation of a shorter window being easier to predict. The significance of terms is disappointing. They seem to say that only the volatility lag was significant, with the exception of *tendies* in the last window which has a t-statistic of -2.103

While my results suggest that price variation is not directly related to the comments on the WSB subreddit, it seems likely that such social media activity has the ability to activate other retail investors for the given cause. However, given that this event had malintent, i.e. to short squeeze institutional investors, it raises important questions for future studies. There was a lot that could have been done differently in this study. Regressing price variation against the ratio of counts was not my idea originally and seems inefficient when I look at it now. Unfortunately, due to my previous regression being done with incorrect/incomplete data, this was more of a last-minute decision; however, with the correct data. Perhaps including volume and other price indicators aside from the Yang-Zhang could have helped explain the next day's volatility.

4. Tables

Table 1: Regression results using a 3-day Yang-Zhang volatility window.

Dep. Variable:	V_GME	R-squared:	0.936
Model:	OLS	Adj. R-squared:	0.932
Method:	Least Squares	F-statistic:	244.5
Date:	Sun, 01 Aug 2021	Prob (F-statistic):	1.71e-48
Time:	17:57:57	Log-Likelihood:	145.74
No. Observations:	90	AIC:	-279.5
Df Residuals:	84	BIC:	-264.5
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0002	0.041	-0.004	0.997	-0.081	0.081
V_LAG	0.9619	0.039	24.503	0.000	0.884	1.040
W_GME	0.0094	0.021	0.456	0.649	-0.032	0.051
W_YOLO	0.0256	0.020	1.289	0.201	-0.014	0.065
W_STONKS	-0.0084	0.019	-0.434	0.665	-0.047	0.030
W_TENDIES	-0.0277	0.017	-1.653	0.102	-0.061	0.006

Omnibus:	8.613	Durbin-Watson:	1.134
Prob(Omnibus):	0.013	Jarque-Bera (JB):	18.646
Skew:	-0.084	Prob(JB):	8.94e-05
Kurtosis:	5.223	Cond. No.	27.2

Table 2: Regression results using a 7-day Yang-Zhang volatility window.

Dep. Variable:	V_GME	R-squared:	0.845
Model:	OLS	Adj. R-squared:	0.836
Method:	Least Squares	F-statistic:	91.79
Date:	Sun, 01 Aug 2021	Prob (F-statistic):	1.53e-32
Time:	17:57:11	Log-Likelihood:	88.380
No. Observations:	90	AIC:	-164.8
Df Residuals:	84	BIC:	-149.8
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.1433	0.070	2.043	0.044	0.004	0.283
V_LAG	0.8424	0.057	14.801	0.000	0.729	0.956
W_GME	0.0531	0.038	1.391	0.168	-0.023	0.129
W_YOLO	0.0577	0.036	1.582	0.117	-0.015	0.130
W_STONKS	0.0163	0.036	0.452	0.653	-0.055	0.088
W_TENDIES	-0.0330	0.032	-1.040	0.301	-0.096	0.030

Omnibus:	13.643	Durbin-Watson:	1.289
Prob(Omnibus):	0.001	Jarque-Bera (JB):	34.385
Skew:	0.394	Prob(JB):	3.41e-08
Kurtosis:	5.924	Cond. No.	22.8

Table 3: Regression results using a 14-day Yang-Zhang volatility window.

Dep. Variable:	V_GME	R-squared:	0.973
Model:	OLS	Adj. R-squared:	0.971
Method:	Least Squares	F-statistic:	599.5
Date:	Sun, 01 Aug 2021	Prob (F-statistic):	4.02e-64
Time:	17:58:25	Log-Likelihood:	197.52
No. Observations:	90	AIC:	-383.0
Df Residuals:	84	BIC:	-368.0
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0616	0.020	-3.055	0.003	-0.102	-0.022
V_LAG	1.0262	0.023	44.353	0.000	0.980	1.072
W_GME	-0.0219	0.011	-1.931	0.057	-0.045	0.001
W_YOLO	-0.0067	0.011	-0.591	0.556	-0.029	0.016
W_STONKS	-0.0041	0.011	-0.384	0.702	-0.025	0.017
W_TENDIES	-0.0177	0.010	-1.849	0.068	-0.037	0.001

Omnibus:	16.221	Durbin-Watson:	1.178
Prob(Omnibus):	0.000	Jarque-Bera (JB):	59.867
Skew:	-0.316	Prob(JB):	1.00e-13
Kurtosis:	6.945	Cond. No.	25.7

Table 4: Regression results using a 21-day Yang-Zhang volatility window.

Dep. Variable:	V_GME	R-squared:	0.984
Model:	OLS	Adj. R-squared:	0.983
Method:	Least Squares	F-statistic:	1011.
Date:	Sun, 01 Aug 2021	Prob (F-statistic):	1.91e-73
Time:	17:58:44	Log-Likelihood:	230.81
No. Observations:	90	AIC:	-449.6
Df Residuals:	84	BIC:	-434.6
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0456	0.012	-3.744	0.000	-0.070	-0.021
V_LAG	1.0141	0.015	65.429	0.000	0.983	1.045
W_GME	-0.0144	0.008	-1.870	0.065	-0.030	0.001
W_YOLO	-0.0139	0.007	-1.916	0.059	-0.028	0.001
W_STONKS	0.0010	0.007	0.130	0.897	-0.014	0.016
W_TENDIES	-0.0141	0.007	-2.103	0.038	-0.028	-0.001

Omnibus:	11.044	Durbin-Watson:	1.363
Prob(Omnibus):	0.004	Jarque-Bera (JB):	31.455
Skew:	-0.033	Prob(JB):	1.48e-07
Kurtosis:	5.895	Cond. No.	23.0

5. Figures

Figure 1: Plots the two inputs for the WSB interest indicator, term occurrences for the WSB subreddit and all of Reddit.

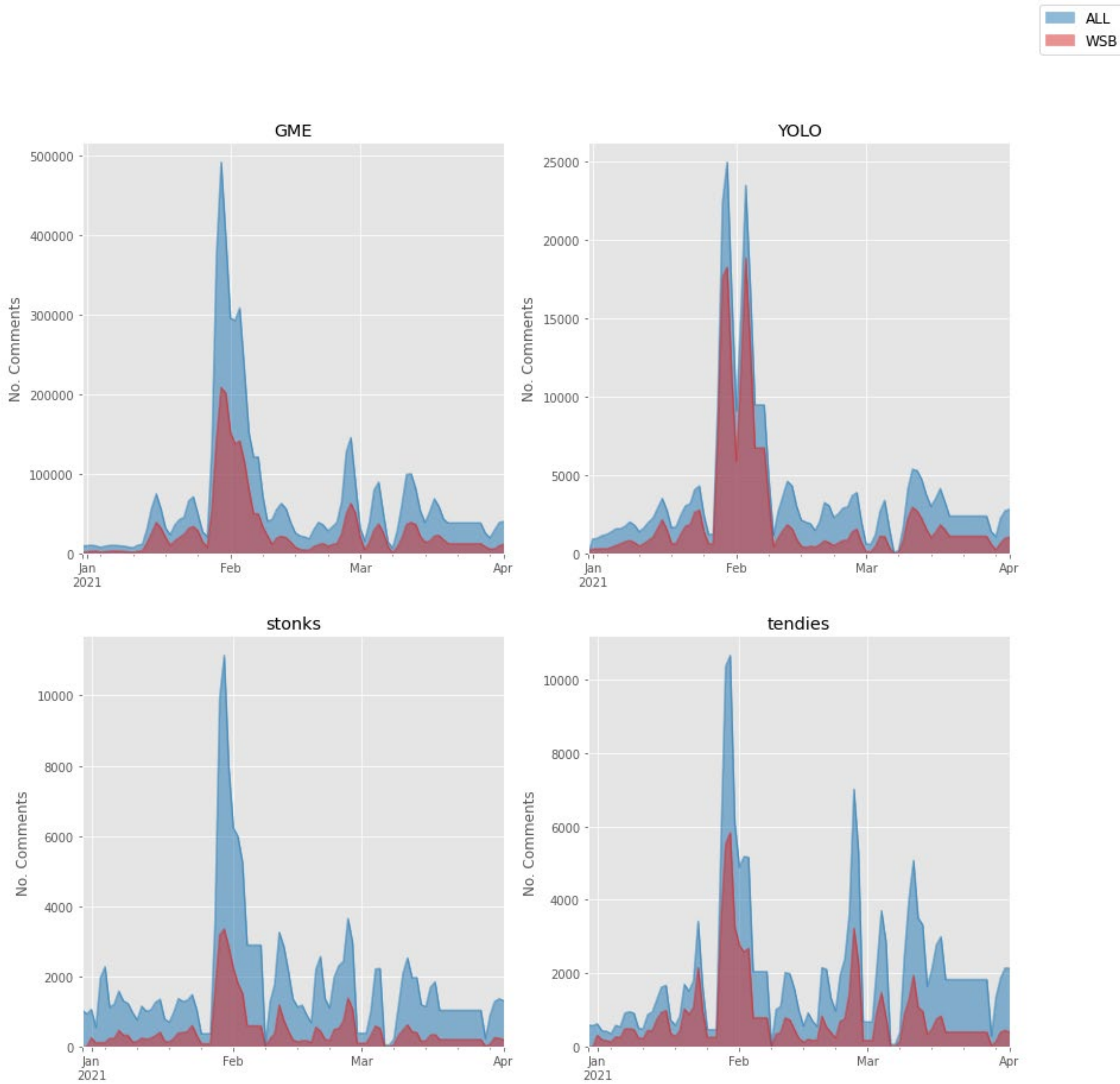
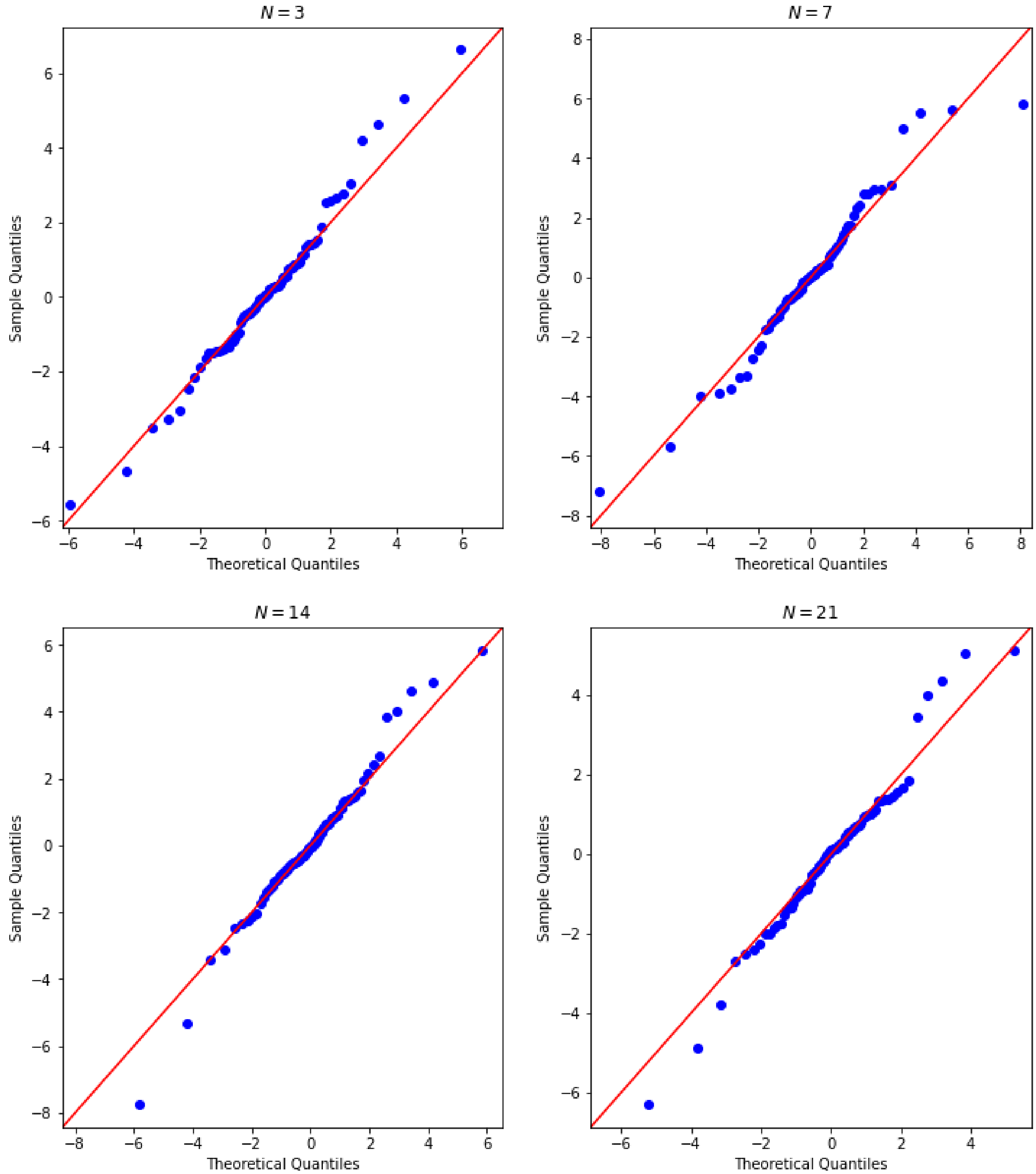


Figure 2: Q-Q plot of the quantiles of residuals versus the quantiles of the t-distribution.



6. References

Yang, D. and Zhang, Q. (2000). Drift-independent volatility estimation based on high, low, open, and close prices. The Journal of Business, 73(3):477–492.

Rogers, L. C. G. and Satchell, S. E. (1991). Estimating variance from high, low, and closing prices. *The Annals of Applied Probability*, pages 504–512.

Rui X (2015) Google search volume index: predicting returns, volatility, and trading volume of tech stocks. Duke University Economics Department. https://sites.duke.edu/djepapers/files/2016/10/xurui-dje.original.pdf.