Empirical Research on Skewness in Individual U.S. Stock Log Returns at Different Investment Time Horizon*

Daehyun Kim†

December 01, 2017

Abstract

This paper assumes that a stock log return is an independent and identically distributed (i.i.d.) random variable and studies the change in the skewness of its distribution as the investment horizon increases. The theoretical expectation is that the absolute value of skewness must become smaller as investment horizon increases, however, the empirical tests on 3,194 firms in the US stock market disprove it. This is either because the stock log returns are not i.i.d. random in general, or the estimator used in the test is not accurate enough to describe the true nature of skewness. The paper concludes without rejecting the null hypothesis and proposes room for improvement on future research.

Keywords: Skewness, Distribution of Stock Log Return, Time Horizon, Investment Horizon.

*This paper is a final version of an undergraduate research paper for Advanced Finance Seminar (FINE 547) in Desautel Faculty of Management, McGill University.

†Undergraduate in McGill University; dae.h.kim@mail.mcgill.ca

Acknowledgement

I sincerely show my gratitude to professor Vihang R. Errunza for his critical mentoring in regards to devising and modeling the theoretical expectations of this research.

Contents

1 Introduction

Papers in finance often treat a stock return as a sequence of independent and identically distributed (i.i.d.) random variables. Some theories assume that the distributions of stock returns generally follow a normal distribution, while other papers disagree and comes up with empirical evidences that stock returns are not normal. The distribution of returns are often described as being asymmetric and having fatter tails. While the shapes of such distributions are diverse and not specifiable, there are still some attempts to approximate them and to reduce error. This paper mainly focuses on the third central moments, also known as the skewness, of the distributions of individual stock returns of the firms present in the current US market. I performed an empirical analysis on 3,194 firms currently[1] present on US Stock Market to determine if increasing the time horizon can improve on such skewness of the individual distributions of returns and potentially reduce error originating from the asymmetry in the future research.

Throughout this paper, one major assumption is that a stock's log return is an i.i.d. random variable, which its distribution is not necessarily normal but rather unperceivable. The theoretical expectation under this premise is that the absolute value of the skewness of such i.i.d. random variable must decrease, as the investment horizon increases. Eriksson(2008) comes up with the formula of skewness of i.i.d. random variable, expressed as a following equation:

[1]The data is lastly updated on November 09, 2017.

$$\text{Skew}(r_{IH=n}) = \text{Skew}(r_1 + r_2 + r_3 + \cdots + r_{n-1} + r_n) = \frac{1}{\sqrt{n}} \cdot \text{Skew}(r)$$

where:

n = investment horizon in trading days

$r_{IH=n}$= stock log return with n days of investment horizon

r_t= stock log return at each trading day t with one day of investment horizon

(= i.i.d. random)

Therefore: if n increases, the absolute value of the skewness must become smaller.

The equation is simple and elegant, by the virtue of the additive nature of log returns when investment horizon increases. For example, the log returns for investment horizon of 2 days is the summation of the log returns of two consecutive days.

$$\text{log return for two days} = \text{log returns for day}_1 \qquad +\text{log return for day}_2$$

which is equivalent to:

$$\ln(P3) - \ln(P1) = \ln(P3) - \ln(P2) \qquad + \ln(P2) - \ln(P1)$$

which is equivalent to:

$$\ln\left(\frac{P3}{P1}\right) = \ln\left(\frac{P3}{P2}\right) + \ln\left(\frac{P2}{P1}\right)$$

which is equivalent to:

$$r_{IH=2} = r_2 + r_1$$

where:

Pt = the adjusted stock price at day$_t$

Same logic applies to log returns with investment horizon of 3 days or longer.

2 A brief overview of the key definitions

Logarithmic return, or log return, of a stock for a certain period of time n days, is defined as

$$\text{Log return}_{(IH=n)} = \ln\left(\frac{P_n}{P_0}\right) = \ln(P_n) - \ln(P_0)$$

, where P_0 is the price of a stock on a certain day and P_n is that of the same stock after n days. In the dataset of this paper, adjusted closing prices are used for P_0 and P_n, which are immune to dividend payoff and stock split.

Time horizon, or investment horizon (IH), is the total length of time (in the unit of trading day, in this paper) that an investor expects to hold a stock before evaluating its return. In the above equation, this is represented by n, the number of days between the observation of P_2 and P_1. The symbol ($IH = n$) has the meaning that the investment horizon is n days long.

2.1 Hypothesis

2.1.1 Null hypothesis

$$\text{Skew}(r_{IH=n}) = \text{Skew}(r_{IH=m})$$

where: $m > n$

2.1.2 Alternative hypothesis

$$\text{Skew}(r_{IH=n}) > \text{Skew}(r_{IH=m})$$

where: $m > n$

Interpretation: the skewness of a log return of a stock at shorter investment horizon is larger than the skewness of that of the same stock at longer investment horizon.

3 Design

The test proceeds with two steps. First step is to explore the whole dataset without any restrictions in terms of the intuitive nature of financial market, such as considering 'market volatility' or 'maximum investment horizon'. It is the second step those considerations are taken into account; the test is given certain restrictions and the results is compared with the first step. See each subsections to read the details about the differences between two steps.

3.1 Data description

'Wiki EOD Stock Prices'[2] is a public domain dataset of end-of-day stock prices, dividends and splits for around 3,000 US companies. As it being a public domain dataset curated by an online community, the data might not be guaranteed to be accurate. Nonetheless, I choose this as the main dataset of this research due to its advantage of being available for free, compared with other alternative proprietary datasets such as CRSP Stock Database[3] [4]. Precisely, the dataset includes 3,194 firms listed in the U.S. Stock market as of November 10, 2017. The starting date of stock price is different among firms, as their establishment dates differ; the earliest date starts from January 2nd, 1962.

[2] from Quandl (https://www.quandl.com/databases/WIKIP)

[3] The Center for Research in Security Prices (http://www.crsp.com/products/research-products/crsp-us-stock-databses)

[4] As this paper is not exclusive on a certain dataset, it can be handily replaced with other authoritative dataset at anytime in future improvement with the presence of funding, sponsorship or academic access.

3.2 Data processing

MATLAB programming language[5] is used for the whole process of the analysis, since its strengths in parallel computing and matrix operation suits the goal of this research[6]. The dataset is a long form of panel data (15,153,834 rows x 14 columns) and each company comes with its own unique ticker. Extraction of each firm's data is necessary, as the i.i.d. assumption of a stock return only holds within the same entity. I generated 3,194 sub-datasets, where each sub-dataset containing stock price information about each firm, from the original data pool.

As the dataset is large, it requires extensive hours of computation for testing all existing firms without compensating too much for the number of investment horizons tested. An efficient way to explore the dataset is to raise the investment horizon by the power of two (i.e. daily, 2-day, 4-day, 8-day, . . . , 256-day, 512-day, 1024-day, . . .). Notice that, by chance, 256-day is a good approximation for an investment horizon of a year (252 trading days). This is efficient in the way that, given a vector of log returns with investment horizon of n, a new vector of log returns with investment horizon of $2n$ can be easily achieved by adding up the odd and the even entities of the old vector. In MATLAB, this is expressed as:

```
% ret is a vector of log returns
ret = ret(1:2:end-1) + ret(2:2:end);
skewness(ret);
```

Again, this is possible by the virtue of the additive nature of log returns. As I iterate through the above line repeatly, new vectors of log returns for the next two-fold increased investment horizon and their skewnesses are achieved. One major flaw of this approach is that the sample size decreases by half for each iteration (i.e. the length of the new 'ret'

[5]MATLAB (https://www.mathworks.com/products/matlab.html)

[6]There is no reason to stick to MATLAB and the code can be simply ported to other programming languages.

vector is $\frac{1}{2}$ of the old 'ret' vector) and the estimates for longer investment horizon would become inaccurate; there may be a statistical bias or inconsistency coming from different sample sizes. See 'Limitations' section to read the details about this issue.

3.3 Calculation of skewness for n days of investment horizon

3.3.1 The first step: testing on full data without investment horizon limit

As seen from the above code example, MATLAB's builtin skewness function is used to estimate the skewness of log returns. By definition of skewness, at least more than 2 samples are required to generate non-zero estimates. Otherwise the estimate is either zero or NaN[7].

For this trial, no upper limit for investment horizon is set. In most cases this does not intuitively make sense as the business environment and market conditions change over time, and the firm itself also changes in its quality; a firm today and the same firm in the past, i.e. 30 years ago, cannot be considered as the same entity anymore, violating the i.i.d. assumption. However, it is difficult to come up with a standardized measure to control for these unquantifiable characteristics. So I tested the data as-is for the first step, set regulations in the second step and then compared the results.

Figure 2 shows the change in skewness as investment horizon increases by a factor of two. It seems the skewness estimates converge to zero following the increase in $\log_2$ days of investment horizon, which is the x-axis. However, this is not true if I analyze the numerical results, putting aside visual effects. When calculating the proportion of firms which have larger skewness estimates at 1-day investment horizon than their skewness estimates at longest available skewness estimates (this might vary across different firms depending on how many log returns data are available)[8],

[7] a special representation of Not-a-Number value in MATLAB

[8] For the first step trial with full data, minimum of 3 samples are the requirement to produce an estimate

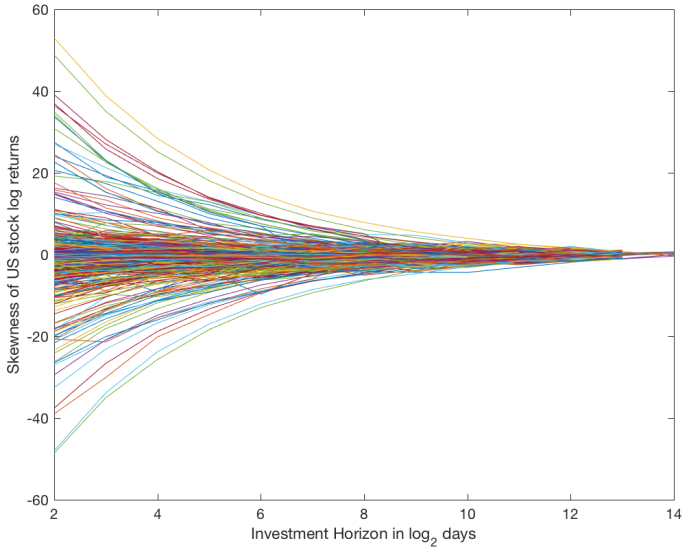

Figure 1: Skewness estimates of log returns of 3,194 U.S. firms

```
1629 / 3193 = 5.101785e+01
```

Only 51 per cent of the firms out of 3,193 firms[9] turned out to have a larger absolute value of the skewness at 1-day investment horizon. This is a disappointing result. I extended the proportion analysis further, and now I averaged out the absolute values of two skewness estimates at 1-day and 2-day investment horizons of each firm and compared it with the mean of absolute values of two skewness estimates at the longest, and the second longest investment horizons.

```
1490 / 3193 = 4.666458e+01
```

The result was even more disappointing, since only 46.7 per cent of the total 3,193 firms had a larger mean of absolute values of the skewness estimates at shorter investment horizon.

for skewness. In order to produce a single skewness estimate for 8-log_2 days, which is the same as 256-day investment horizon, a firm must have three or more records of stock returns.

[9]one firm with ticker 'TPR' only had one skewness estimate, not being eligible for the proportion test

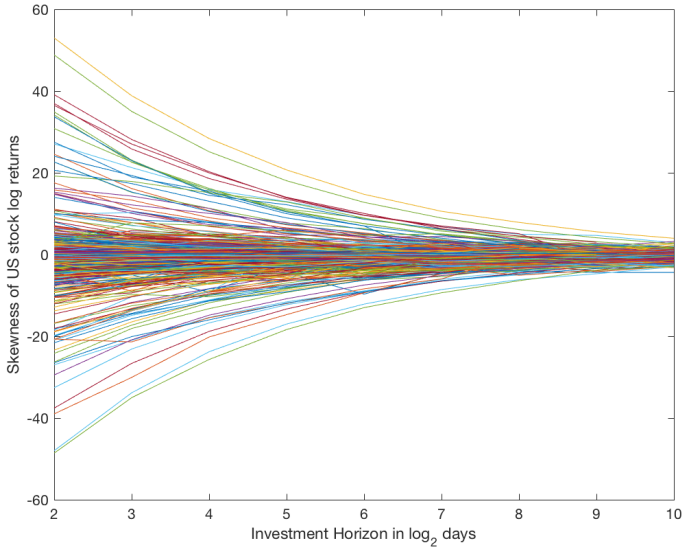

Figure 2: Skewness estimates of log returns of 3,194 U.S. firms, with minimum sample size requirement:n¿=10, max investment horizon: 1024 days

I extended the analysis for three and four skewness estimates with the same algorithms,

```
with n=3 average : 1431 / 3188 = 4.488708e+01
with n=4 average :  371 / 3169 = 4.326286e+01
```

and the results were still not noteworthy. In fact, these results are the opposite of what theoretical expectation predicts.

3.3.2 The second step: testing on modified data with minimum sample size and maximum investment horizon limit

Since the estimates with around 3 samples are inaccurate, I introduce a minimum requirement for sample size of 10. There is a must be a compensation in sample size with its accuracy; if the sample size is too large such as 30, a firm needs to have at least thirty years of stock price history to generate a single skewness estimate at yearly investment

horizon, which is implausible for most cases. Although sample size of 30 is often considered as the most appropriate minimum sample size for being the general rule of thumb to apply Central Limit Theorem, average firms do not have enough stock price history to meet the minimum requirement. Therefore I decided to find a reasonable, yet not overly compensating sample size requirement, which is set to 10.

In addition, I set a maximum investment horizon limit of 10 $\log_2$ days, in order to avoid violating the i.i.d. asssumption of a stock return by investing long enough to change the attributes of the invested firm. I assume a practical max length of time an investor would decide to put her fortune in without revising the firm's nature is around 3 years, which approximates to 1024 days, or 10 $\log_2$ days. Here are the results applying the same proportional tests as in the first step.

```
firms with abs(skew(1)) >= abs(skew(end)) : 1561 / 3193 = 4.888819e+01
n = 2 average skew                        : 1425 / 3189 = 4.468485e+01
n = 3 average skew                        : 1382 / 3174 = 4.354127e+01
n = 4 average skew                        : 1194 / 2846 = 4.195362e+01
```

The results are all below 50 per cent, which I cannot reject the null hypothesis.

4 Conclusion

The visual shapes of two plots creates an illusion that skewness estimates converge to zero as the investment horizon increases, however, the proportion analysis reveals that skewness estimates does not become smaller with increased investment horizons. The delusion comes from some outlier firms with excessive absolute value of skewness estimates at short investment horizons; if those outliers are removed from the plots, a figure of a thick horizontal column formed by a dense population of oscillating lines remain, indicating that

the trend is unpredictable.

One possible reason for these test results is that stock returns might not be i.i.d. distributed. Since a stock return is affected by diverse factors which change significantly over time, the distribution of the random variable might not be identical. Stock returns might not be independent to each other, there might present fixed time factors which make stock returns autocorrelated.

Another possibility is that skewness of stock returns are actually i.i.d. and they do reject the null hypothesis and follow the alternative hypothesis, however the estimators used to approximate the true value of skewness are not accurate enough. Either the sample size for estimates were not large enough (especially with those at longer investment horizons), or there is a bias caused by different sample sizes among skewness estimates.

5 Limitations and proposals for improvement

There might be some periods of a market where firms' stock returns act as i.i.d. random variable and other periods where i.i.d. conditions cannot hold. e.g. in highly volatile market condition. This test is not adjusted for the external market conditions. For future research, skewness estimates will only be derived from a special sample data under specific market conditions, e.g. from a calm period of low market volatility.

6 Appendix

6.1 MATLAB code

The code is uploaded on github.[10] Average processing time (macbook air):

- To prepare the prerequisite mat-files (firmsWithId.mat): 120 minutes
- To load mat-files : 20 minutes
- Running time after optimization: 30 seconds

6.2 Data Source

Wiki EOD Stock Prices' downloaded from Quandl on November 09, 2017[11]

7 Bibliography

Eriksson, M. (2008). A simulation method for skewness correction, 29-30.

Peiro, A. (2001). Skewness in individual stocks at different investment horizons Quantitative Finance, volume 2, 2002, 139-146.

[10] https://github.com/dqgthb/skewnessInvestmentHorizon
[11] (https://www.quandl.com/databases/WIKIP)