Quantitative Research Interview Assignment

Richfox Capital

Volatility Analysis and Prediction

Kelvin Brinham

GitHub Repository Link

Date: February 12, 2023

Contents

1 Objectives

In summary, this project investigates the liquidity and volatility of four fictional stocks using intraday (volume and price) trading data.

1. Investigate the relationship between the volatility of each stock and its liquidity.
2. Predict stock volatility (in annualized percentage return) for the week following the end of the period of trade data.

2 Data Cleaning

2.1 Data Cleaning

The data provided for each of the four stocks is a list of trades which detail the time, price and volume traded. Given the frequency of data provided, it is unsurprising that errors are present. We remove these errors according to the following procedure:

1. Remove duplicate time entries (keeping the first occurrence)
2. Remove invalid/missing data (e.g. non-numeric data or negative prices/volumes)
3. Remove data outside market hours (08:00 - 16:30 for stocks A and B; 08:00 - 16:00 for stocks C and D)

The statistics relating to how much data was removed are outlined in table 1. We chose to remove invalid/missing data rather than replace it because the data set is large.

Stock	Unclean Size	Repeated Entries	Rows with Missing/Invalid Data
A	182288	0	51
B	92238	0	91
C	115998	0	104
D	69588	0	188

Table 1: Summary of data removed during the cleaning process.

3 Data Processing and Calculating Variables of Interest

The raw data is high frequency and thus exhibits significant microstructure noise. In addition, the time series' are asynchronous. Therefore, we transform the data into 5-minute intervals. Price is taken to be the price immediately before the timestamp, except for the 8:00 value which takes the first price value for each trading day. This method is known as the previous tick method which is preferable to linear interpolation in high-frequency economic time series because linear interpolation can introduce significant bias (e.g. Gençay et al. 2001 Barucci & Reno 2002 Kanatani et al. 2004). Volume is taken to be the summed volume for each 5-minute period.

3.1 Realised Volatility

Volatility can be described as a measure of stock price variation. Volatility is commonly encoded in the Realised volatility (RV). It is calculated from (log) price returns,

$$r(t) = ln\left(\frac{P(t)}{P(t-1)}\right) \tag{1}$$

where $P(t)$ is the price at the timestamp t and $P(t-1)$ is the price at the previous timestamp (5 minutes earlier).

RV is defined as

$$RV = \sum_{t}^{T} r(t)^2 \tag{2}$$

where we choose T to be 1 trading day. This equates to 101 5-minute intervals for stocks A and B and 95 intervals for stocks C and D for which RV is calculated[1]. Further, our definition of RV is daily. Volume is taken to be the sum of all volume data for each trading day. To ensure that 5-minute return calculations did not spread overnight we chunked the data into daily data frames for processing.

We note that there was one instance in which there was no data within a 5-minute interval. In this case, we replaced the return data with the mean from the rest of the day.

3.2 Summary of Data

The final step in processing the data was normalised is using a z-score. Table 2 outlines the cleaned dataset which now contains two daily variables, RV and volume.

Stock	Time Period	Size (Days)	5-minute Periods Per Day
A	01/03/2017 - 18/08/2017	121	95
B	01/03/2017 - 18/08/2017	121	95
C	01/03/2017 - 18/08/2017	121	101
D	01/03/2017 - 18/08/2017	121	101

Table 2: Summary of data after cleaning and processing.

4 Relationship between Volatility and Liquidity (Task 1)

Volume is commonly used as a measure of liquidity (e.g. Cheriyan & Lazar 2019, Le & Gregoriou 2020), in our case daily volume (just volume hereafter) is considered. As mentioned above, our

[1]There are 95 and 101 5-minute intervals respectively rather than 96 and 102 because we have no 5-minute return for 08:00 (see equation 1).

chosen measure of volatility is daily RV (just RV hereafter) (see §3.1 for definition).

4.1 Pearson's Correlation

To examine the basic relationship between RV and volume we generate a correlation matrix. As we can see in Table 3, there is no significant correlation between volume and RV. This is in sharp contrast with the majority of studies which have conclusions summarised in Shrestha (2017).

Stock	volume
A: RV	0.14579
B: RV	0.05474
C: RV	-0.03678
D: RV	-0.11301

Table 3: Pearson's correlation matrix for each stock using 5-minute returns to calculate RV.

This lack of correlation may be caused by significant noise in the 5-minute returns used to calculate RV. To reduce this noise we repeated the analysis by calculating RV from the hourly log return. We calculated the hourly log return from the mean hourly price which is the hourly mean of the 5-minute prices produced from the cleaning process. From this new process we produce a correlation matrix (shown in Table 4) indicating a positive correlation between RV and volume for stocks A and B and no significant correlation for stocks C and D. We also include a couple of popular stocks/ETFs for comparison[2]. We also note that stock D had a sudden crash (losing ∼90% of its closing price value in a single trading day) between 19/05/2017 - 22/05/2017.

Stock	volume
A: RV	0.26157
B: RV	0.30994
C: RV	-0.00078
D: RV	-0.09765
AAPL: RV	0.57790
S&P500: RV	0.20358

Table 4: Pearson's correlation matrix for each stock/ETF using hourly returns to calculate RV. The AAPL and S&P500 hourly data spanned from 16/05/2017 - 06/12/2017 and 17/05/2017 - 06/12/2017 respectively.

[2]We examine these popular stocks/ETFs in the same way as for stocks A-D but using the provided hourly prices.

5 Predicting Return Using an ARIMA Model (Task 2)

Autoregressive integrated moving average (ARIMA) models are commonly used for financial time series prediction and can be valuable for short-term predictions (e.g. Ariyo et al. 2014). In particular, ARIMA is used to analyse (generally) stationary time series.

ARIMA models use regression of previous (lagged) values and have a linear combination of error terms which depend on an error on lagged values. ARIMA accepts three terms which are briefly summarised as follows:

- p: Number of lags included in the autoregressive (AR) model
- d: Degree of differencing between subsequent data points
- q: Order of the moving average (MA) model

A common metric for financial ARIMA models is the daily return which is typically calculated from daily closing prices. We take our daily closing prices as the final traded price for each day. A summary of this data, as well as the ARIMA parameters chosen, are outlined in Table 5. A brief discussion of why we chose these parameters is in § 5.1 - 5.3.

Stock	Days	Mean	Std.	p	d	q
A	121	202.53	107.00	1	0	1
B	121	948.91	372.79	1	0	1
C	121	19.68	1.56	1	0	1
D	121	6982.06	5898.65	10	0	12

Table 5: Summary of daily closing price data.

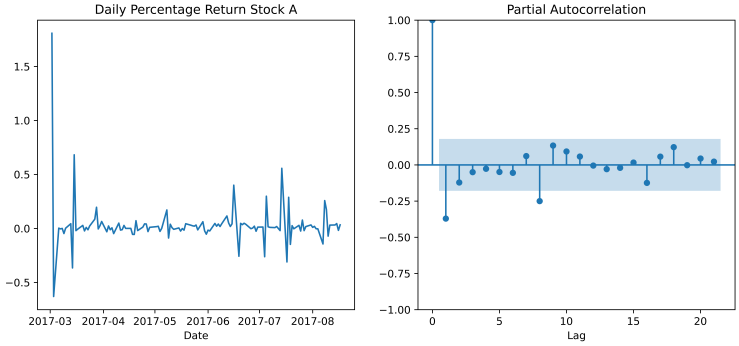
5.1 Stationarity (Choosing d Term)

Intuition suggests return is stationary because it is the first difference of daily closing prices. We can check the stationarity of our 120 percentage daily returns by performing an Augmented Dickey Fuller (ADF) test. Table 6 indicates our daily returns are stationary and that we should pick $d = 0$ for our ARIMA model.

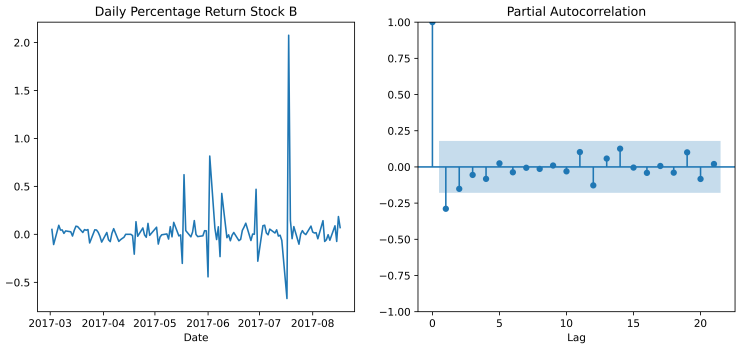
5.2 Autocorrelation (Choosing p Term)

5.3 Autocorrelation (Choosing q Term)

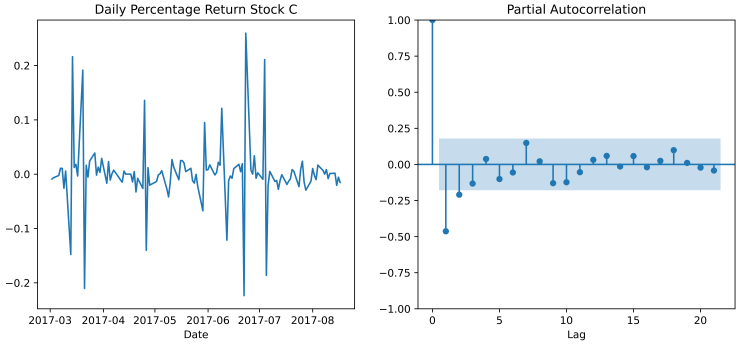
Following a similar procedure to the one in §5.2 we plot autocorrelation plots (ACF) (see Figure 2) to inspect significant autocorrelation at varying lags. We find the most significant autocorrelation occurs for a lag of 1 for stocks A-C and 12 for stock D. Further, we choose $q = 1, 12$ for stocks A-C, and D respectively.



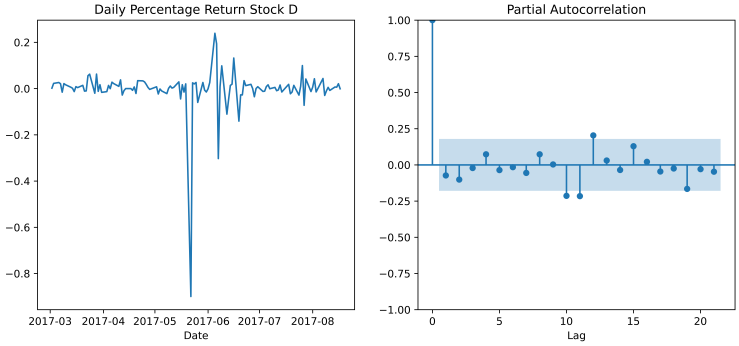
(a) Stock A



(b) Stock B

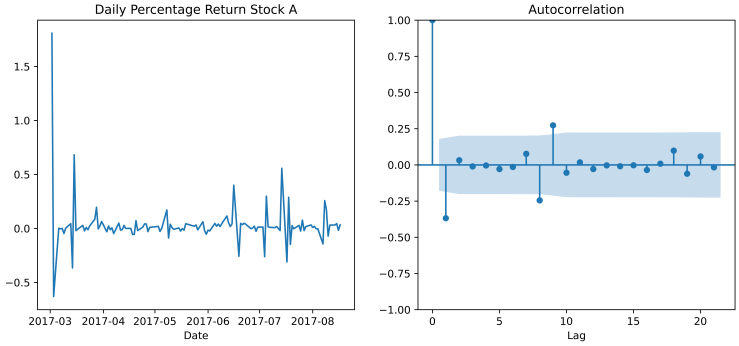


(c) Stock C

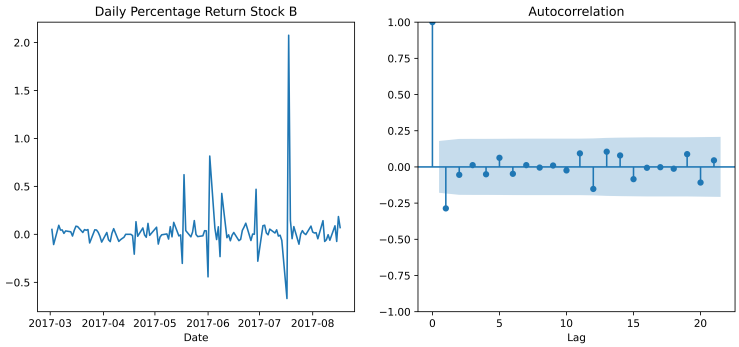


(d) Stock D

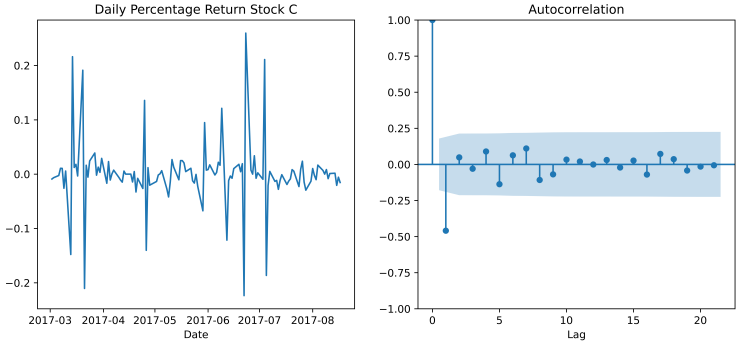
Figure 1: Partial Autocorrelation plots (right) with daily percentage returns (left). The shaded blue region indicates the 95% confidence interval.



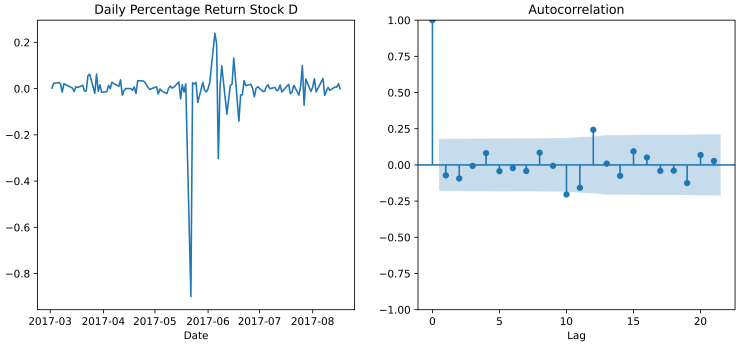
(a) Stock A



(b) Stock B



(c) Stock C



(d) Stock D

Figure 2: Autocorrelation plots (right) with daily percentage returns (left). The shaded blue region indicates the 95% confidence interval.

Stock	t-Statistic	Prob.	1% Critical Value
A	-28.48521	0.0	-3.48653
B	-9.99237	0.0	-3.48702
C	-11.28712	0.0	-3.48702
D	-11.63021	0.0	-3.48653

Table 6: Summary of daily closing price data. The p-values are given to 1 decimal place because they are extremely close to 0, We have included 1% values for comparison.

5.4 Return Prediction Results

We use an ARIMA model to predict the (annualised) percentage return in the week following the data for each stock. The results are outlined in Table 7. There are many downfalls to our method. For instance, ARIMA models require equally spaced time data, but our data runs over non-trading days. To overcome this we have re-indexed the data to effectively ignore non-trading days. In other words, the model takes the return data as being day-to-day for the entire 6 months which ignores the effects of breaks between trading days.

NOTE: These estimations are clearly inaccurate with huge errors (due to a handful of large daily returns) and evidence of significant overfitting. We have included them as part of the process, given the lack of time for further exploration.

Stock	Day	Annualised Return (%)	Upper/Lower 95 %CI
A	1	628	-7583/8839
	2	532	-11350/12414
	3	589	-12337/13515
	4	555	-12718/13828
	5	575	-12818/13968
	6	563	-12871/13998
	7	570	-12879/14020
B	1	-1	-11246/11244
	2	631	-11247/12509
	3	742	-11155/12639
	4	761	-11136/12659
	5	765	-11133/12662
	6	765	-11132/12663
	7	766	-11132/12663
C	1	373	-2245/2991
	2	-16	-3040/3008
	3	24	-3004/3052
	4	20	-3008/3048
	5	20	-3008/3048
	6	20	-3008/3048
	7	20	-3008/3048
D	1	-22	-4182/4138
	2	-379	-4557/3798
	3	149	-4028/4327
	4	801	-3381/4983
	5	-350	-4544/3843
	6	369	-3834/4572
	7	-343	-4546/3859

Table 7: ARIMA annualised predicted daily percentage returns for the 7 days following the data. We define a trading year as 252 days which we use to annualise the daily return predictions.

References

Ariyo A. A., Adewumi A. O., Ayo C. K., 2014, in 2014 UKSim-AMSS 16th international conference on computer modelling and simulation. pp 106–112

Barucci E., Reno R., 2002, Economics Letters, 74, 371

Cheriyan N. K., Lazar D., 2019, International Journal of Economics and Financial Issues, 9, 17

Gençay R., Dacorogna M., Muller U. A., Pictet O., Olsen R., 2001, An introduction to high-frequency finance. Elsevier

Kanatani T., et al., 2004, Economics Bulletin, 3, 1

Le H., Gregoriou A., 2020, Journal of Economic Surveys, 34, 1170

Shrestha S., 2017, PYC Nepal Journal of Management, 10, 40