Estimating the Volatility of Returns

Aditya Raut

November 5th, 2023

Contents

1 Task description

The provided data set contains prices of 6 stocks named a through f, sampled over the course of one year at 1 minute intervals. Our goal in this exercise is to estimate the volatility of these stocks over the month following given data, in the metric of "annualized percent returns". There are a multitude of choices to be made on how the data is processed, the metrics chosen for returns calculation, sub-sampling or aggregating with a different frequency for training, addition of statistical or time series features, choice of model and validation to name a few. We will make use of highly cited papers on the subject and incorporate the observations and conclusions they have derived about datasets depicting stock prices. We make certain assumptions along the way, which will be highlighted in this report.

2 Preliminaries

- For an investment of value V_i which grows to a value V_f over a time period, the *return* or *holding period return* can be calculated as $R = \frac{V_f - V_i}{V_i} = \frac{V_f}{V_i} - 1$. A popular alternative is *logarithmic return*, calculated by $R_{\log} = \log\left(\frac{V_f}{V_i}\right)$.[5]
- Volatility is the degree of variation of trading price over time, usually measured by the standard deviation of logarithmic returns. Annualization of volatility over a period of length T, is multiplication by $\sqrt{N}$, if one year consists of N periods of length T. [6]

3 Data cleaning and exploration

Over the course of 1 year, there are 252 trading days and the trading time horizon is from 9:30 am to 4:00 pm, which are 391 different minute instances. The data set has 98352 rows. Plotting the prices for all 6 stocks brings a problem to attention, that some values for stocks a and d are inconsistent and cause huge fluctuations as seen in figure 1. Upon further inspection, they are rogue observations with values 0 and 1 for the two stocks respectively, and occur with very small frequency. We remove them and fill any missing values for all stocks with **the previous timestamp's value** in the data set. This is our first assumption, that filling data with 1 minute prior data (or last available timestamp) is reasonable.

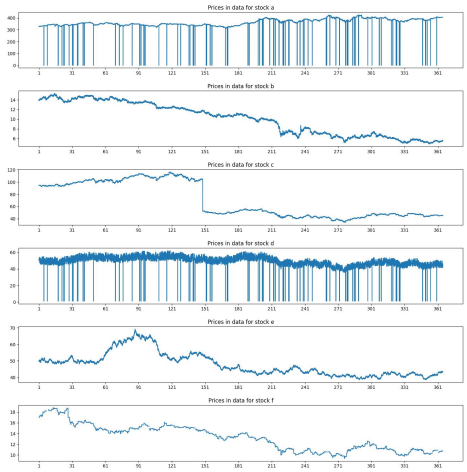

Figure 1: Raw data before any imputation

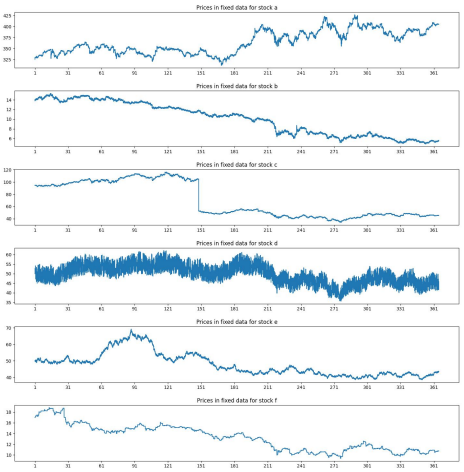

Figure 2: Fixed data after forward filling

The cleaned data more closely resembles usual stock price graphs. We observe that d stock has high daily fluctuations, making it appear much denser than the rest.

3.1 Moving average and scatter plots

We then proceed to plot some moving averages over periods of 1 week to detect time-dependent trends for all the stocks. There are no obvious linear or polynomial trends, with the exception of stock b which follows a linear trend in some part. This is shown in figure 3. We also make scatter plots of pairwise stock prices to find potentially correlated stocks. These initial explorations help us navigate through many options of potential features to be added for predicting this time series, such as time dummies or lag variables. Some scatter plots in figure 4 propose some clustered values, implying some dependence.

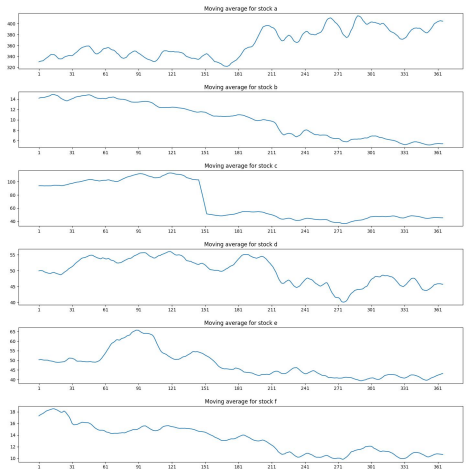

Figure 3: Moving average plot to detect trends

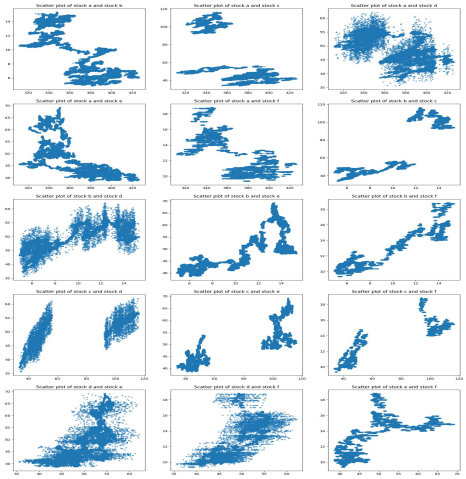

Figure 4: Pairwise scatter plots of all stocks

4 Data modification

4.1 Aggregation to reduce frequency

While surveying related highly cited literature, we came across the results in [4] which state that the direct use of high-frequency data does not improve volatility predictions. This forms the third assumption of our approach, motivating aggregation of data at a day level frequency, significantly reducing the number of rows and hence training time and complexity. We use 'pandas groupby' method for this. Seeing the distribution of daily price ranges, we observe that its variation is very low on most stocks.

4.2 Choice of price to consider

Above aggregation motivates 3 obvious choices for day level aggregation of stock prices, namely the opening, closing and average price. These are defined as the price at 9:30 am, at 4:00 pm and the mean price in that day, respectively. We plot pairwise scatter plots of these 3 prices for all 6 stocks, to observe that they are highly correlated as seen in 5. This leads to our fourth assumption, that we can simply consider one of the three prices moving forward. In our analysis, we stick to using **average daily price**.

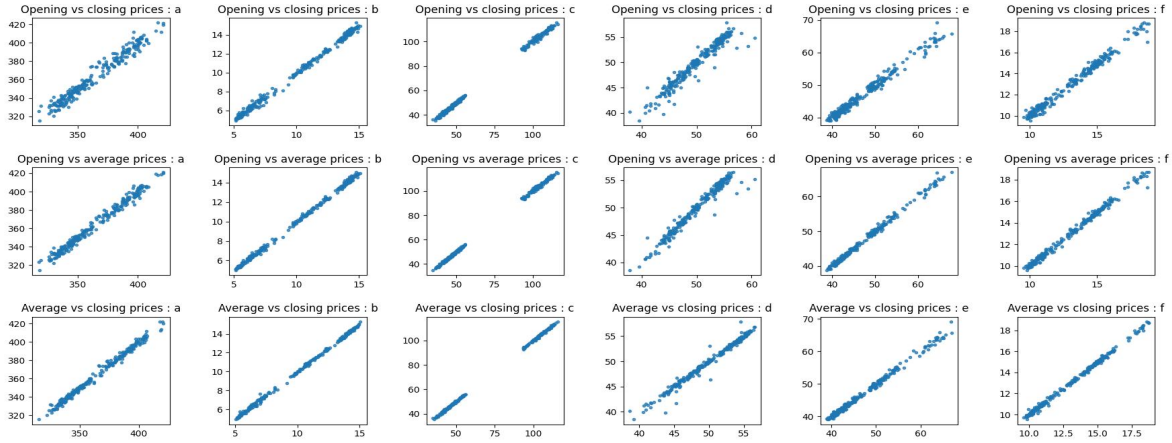

Figure 5: Scatter plots of three natural alternatives for daily price

5 Determining features and target vectors

We calculate returns for the daily stock prices using two definitions in section 2, the holding period and logarithmic returns. It is interesting to note that when x is small, we have $\log(1+x) \approx x$. This suggests that the two return metrics would be closely related to each other as returns scale over small values. This is confirmed by a plot of the two metrics in aggregated daily data as shown in figure 6 where values on the X-axis are trading day counts.

- Since we could not find obvious time-dependent trends in the data, we will use statistical features which have been historically used for volatility prediction.
- Our prediction task needs a lead time of 1 month, since we have to predict volatility 1 month into the future. Thus, use of very recent lag features will not be of much help, as we will not have access to those values for the upcoming month.
- As suggested in [4], data older than 50 days is not very influential for predicting volatility. That is roughly 2 months. So we build a model that uses a previous month's data as features, leaves a lead time of 1 month and predicts the values for the month after that.
- Since calculation of returns is done over 1 month, we only have returns data from 2$^{\text{nd}}$ month onward. Calculation of volatility from returns is done over 1 month by considering a rolling average, so volatility data is available starting month 3. Thus, the first volatility data that can be in target prediction is from day 1 of month 5.

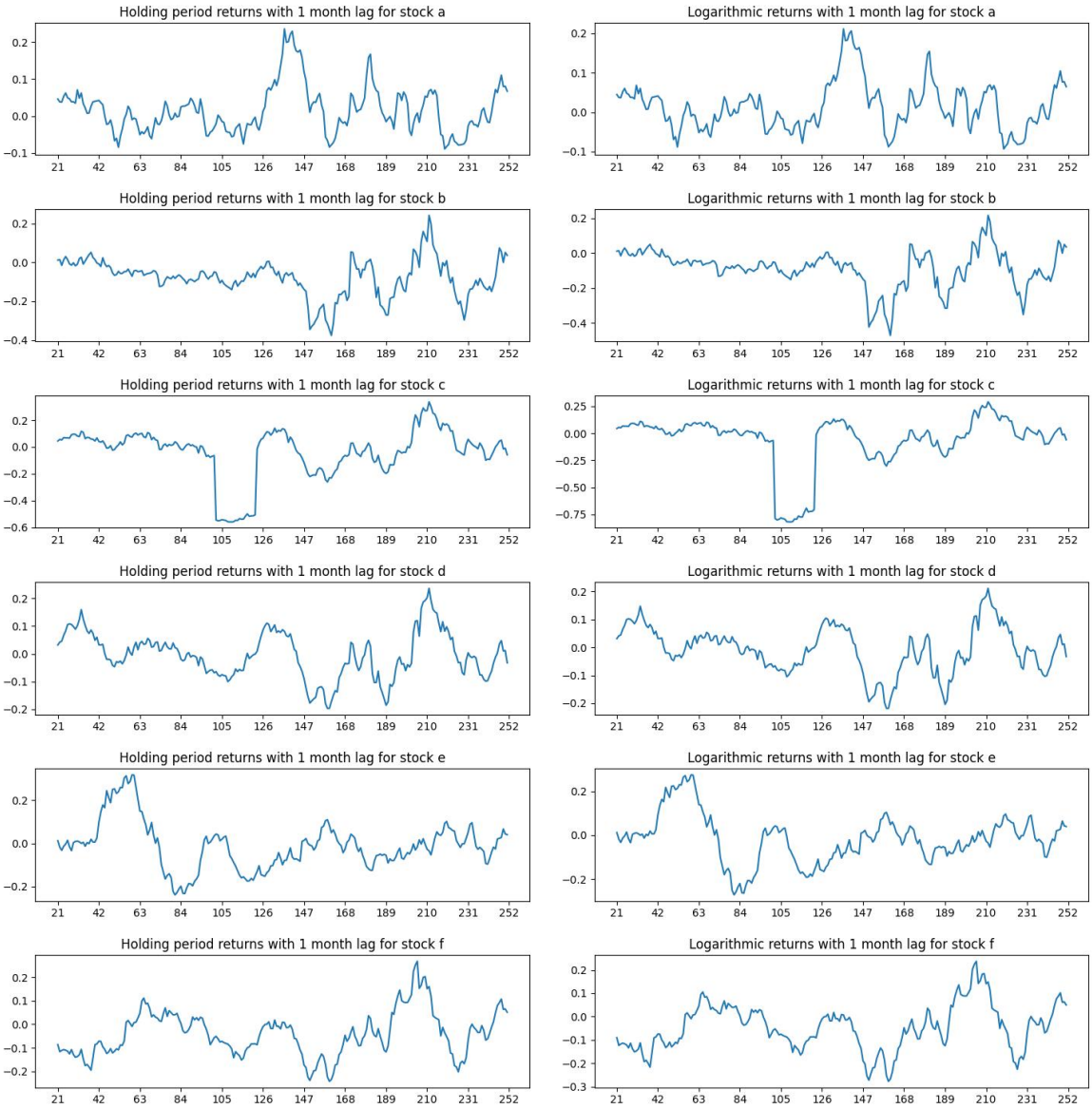

Figure 6: Holding period and logarithmic returns variation over the trading days in a year

Volatility is calculated as month-long rolling standard deviation in the value of logarithmic returns, and is plotted in figure 7 to spot potential trends.

5.1 Feature alternatives

Referring to multiple highly cited papers on volatility predictions, we obtain ideas to use the following potentially good predictors as features for our data.

1. Square of past returns [2]
2. Absolute value of past returns [3]
3. Range of past return values (max-min) [1]
4. Past values of volatility (lag) [4]
5. Daily range of prices [4]
6. Realized daily power, the sum of intra-daily absolute returns [4]

Clearly the 5^{th} and 6^{th} features require data of higher frequency than daily pricing, so we will only look at the first 4 features as potential predictors moving forward. This is one of the **trade-offs** for having reduced data frequency to daily.

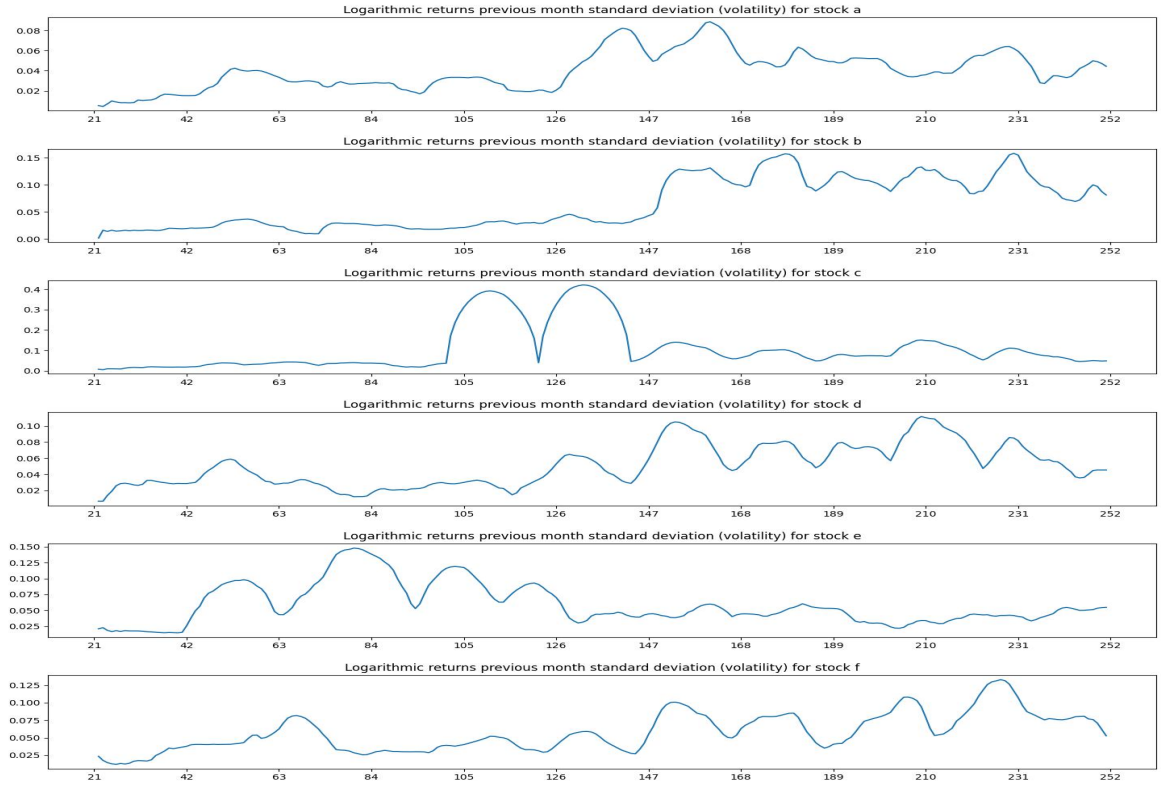

Figure 7: Volatility, the monthly rolling standard deviation of $R_{\log}$

- It will be interesting to see whether each stock's returns are better determined by considering it as a separate time series, or the data completely flattened into one array for all stocks together in training. To save time in selection of features for specific tasks, we write some handy functions.

6 Models, Training, Validation

6.1 Choice of architecture

Since financial data typically has trends detectable by simpler models and for the compelling reason of explainability, we use LinearRegression model for all the training. We perform 20-fold cross validation, training a lot of models to fit patterns in the data. The metric chosen for calculating residuals is Mean Squared Error, as recommended in [4]. Our goal then is to find the best combination of features to model volatility for the next month.

6.2 Selecting the best feature combination

One perk of having easy access to any slices of data with handy functions (mentioned in section 5.1) is that we can train both types of models mentioned earlier (together or separate data for stocks) for a large number of cross validations with mean squared error. Since we have access to 4 features, there are $2^4 - 1 = 15$ different combinations of them potentially being used as training data. **We iterate through all these 15 combinations** of features (named 'Data_presets') during training and save the information for which combination performed the best with a cross entropy validation loss function.

It is observed that using separate training for each stock results in a smaller mean squared error than the best feature combination for combined data. When using separate data for each stock, the smallest mean square error is obtained for the following combination of features (from section 5.1) -

a : {1} b : {4} c : {3,4} d : {2} e : {2,4} f : {3}

7 Results

- After finding out the best selection of features for each stock, we train 6 different linear models that learn volatility information for each stock.
- The relevant feature data for each day of the final month in our data set is then provided as input to the trained model to obtain predictions of volatility for each day of the future month.
- This is different from the exact number "volatility over the next month", which in our definition would simply be the value of volatility value on the last day of the future month.
- We use mean volatility observed over the future month for answering the question of annualized volatility. The answer is obtained by multiplying monthly volatility with $\sqrt{12}$, as discussed in section 2.
- The root mean squared error (fluctuation) serves a similar purpose as standard deviation to conclude confidence intervals, so presumably the usual 68-95-99.7 percent confidence rule for 1-2-3 standard deviation applies here as well.

Stock	Month End Vol.	Month Avg Vol.	M.S. Error	Annualized Vol.	Fluctuation
a	5.04 %	4.62 %	0.000260	16.01 %	5.58 %
b	6.65 %	9.56 %	0.001421	33.13 %	13.06 %
c	17.83 %	10.94 %	0.005972	37.91 %	26.77 %
d	4.16 %	5.93 %	0.000580	20.55 %	8.34 %
e	5.06 %	4.39 %	0.000256	15.21 %	5.54 %
f	6.98 %	7.72 %	0.000684	26.74 %	9.06 %

8 Limitations, future improvements

The fluctuations in result coming from root mean squared error seem to be high, which means the confidence intervals are really wide in this current result. I believe the use of a smarter error metric or better calculation of the error than 95% of training data (coming from 20-fold cross validation) can improve these results and provide a more confident estimation of volatility.

References

[1] S. Alizadeh, M. W. Brandt, and F. X. Diebold. Range-based estimation of stochastic volatility models or exchange rate dynamics are more interesting than you think. 1999.

[2] T. Bollerslev. A conditionally heteroskedastic time series model for speculative prices and rates of return. *The review of economics and statistics*, pages 542–547, 1987.

[3] Z. Ding, C. W. Granger, and R. F. Engle. A long memory property of stock market returns and a new model. *Journal of empirical finance*, 1(1):83–106, 1993.

[4] E. Ghysels, P. Santa-Clara, and R. Valkanov. Predicting volatility: getting the most out of return data sampled at different frequencies. *Journal of Econometrics*, 131(1-2):59–95, 2006.

[5] Wikipedia contributors. Rate of return — Wikipedia, the free encyclopedia, 2023. [Online; accessed 5-November-2023].

[6] Wikipedia contributors. Volatility (finance) — Wikipedia, the free encyclopedia, 2023. [Online; accessed 5-November-2023].

9 Appendix : Explainability of the linear model

As mentioned in section 6.1, the linear model is simple to explain. As observed commonly from the plots here for regression coefficients, more recent data plays a more crucial role in predicting volatility. There also seem to be some periodicity in coefficients along training time horizon.

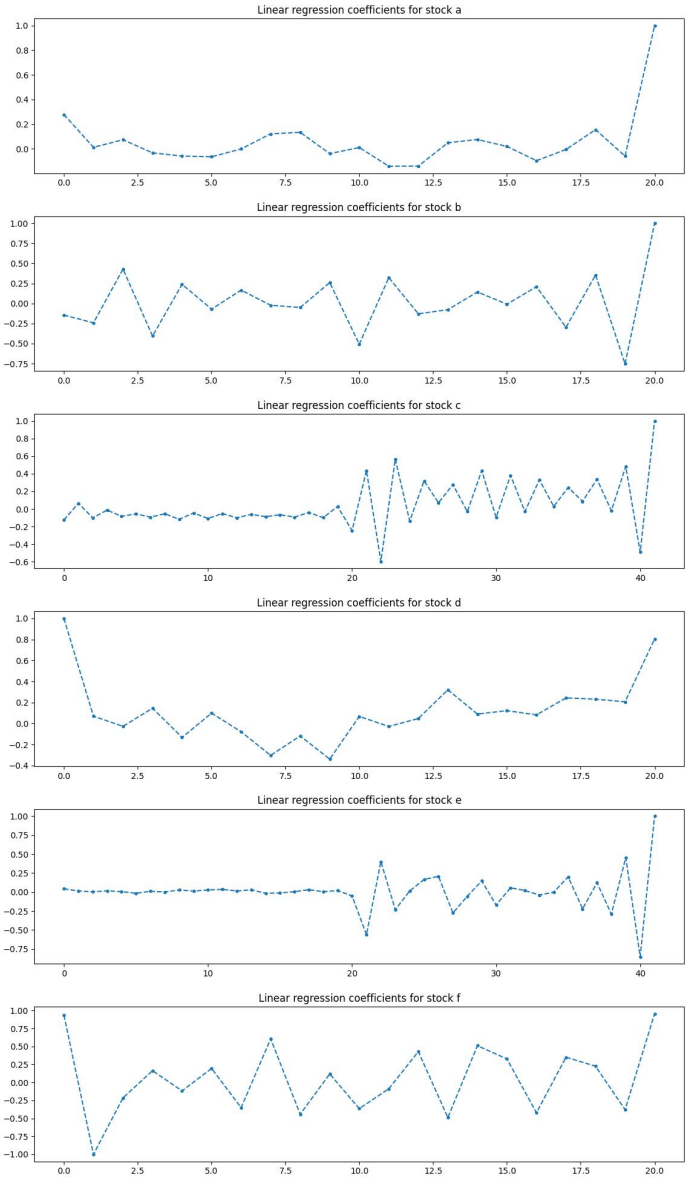

Figure 8: Regression coefficients normalized in the range $(-1, 1)$