Assignment 3

Financial Economics

April 1, 2020

1 Realised variance

Realised variance (RV_t) of an asset with price $p_{i,t}$ on a day t is given by :

$$RV_t = \sum_{i=1}^{n} {r_{i,t}}^2 \tag{1}$$

where $r_{i,t} = log(p_{i,t+1}) - log(p_{i,t})$ [Zamojski, 2020, Corsi, 2009].

With continuous information about the asset prices at any given moment, it can be shown that the RV_t converges towards the true daily variation. However, financial markets are imperfect, even if they are one of the most effective trading spots in the economy. Combined with the fact that RV_t is one of the most precise ex-post measure of volatility, performing calculations on historical data can give a pretty accurate picture of the fluctuations in the prices of assets over time. These can be used for forecasting future price developments, although this might only be done with great caution.

Optimal sampling frequency is often brought up as an issue related to using RV_t. Too frequent sampling can cause the sample to be contaminated by market microstructure noise. At the same time, one must have enough information to be precise. A deciding factor is often the nature of the asset of interest and it's liquidity.

With the provided data, RV_t was calculated and plotted on an optimal, daily 30-minute basis :

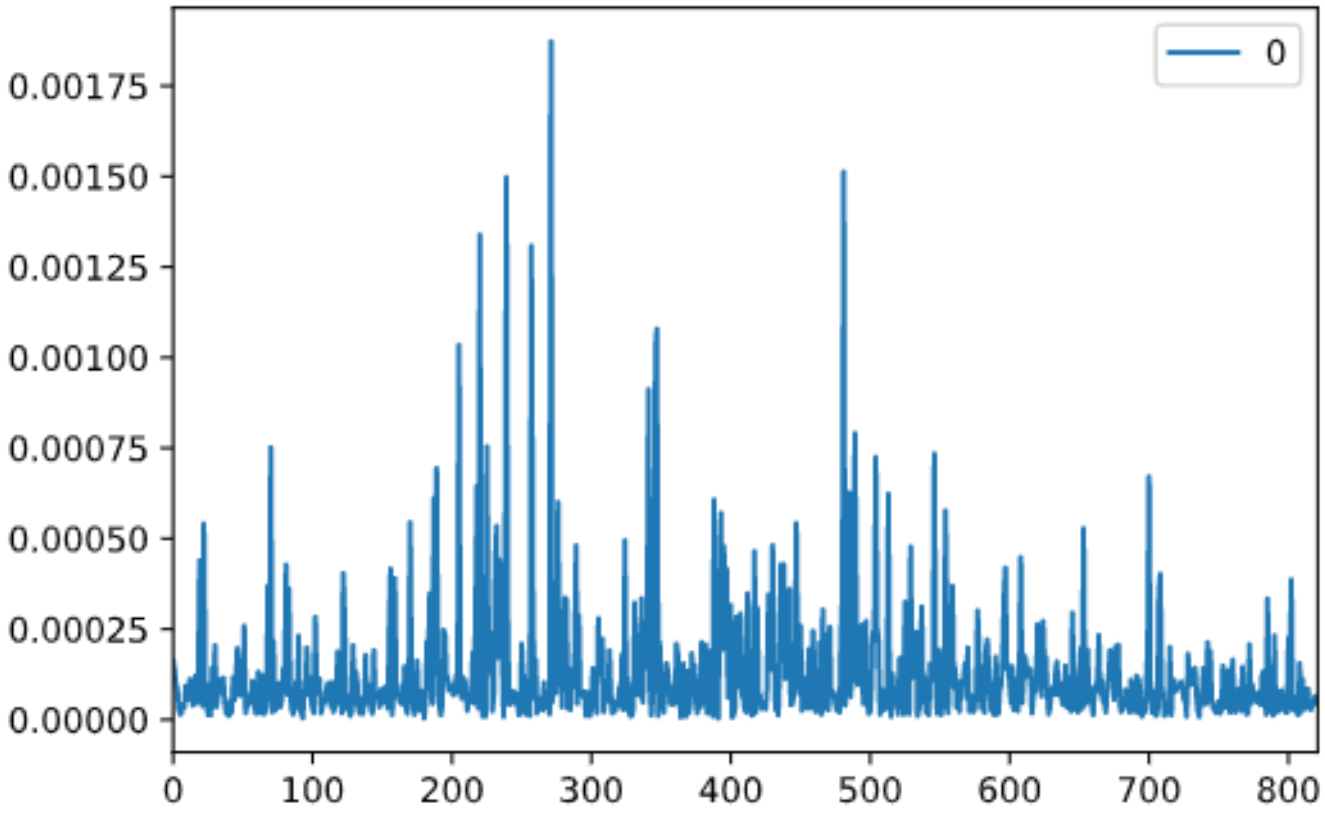

Figure 1: Daily realised variance between January 2007 and June 2010.

On the graph it can be seen that the prices fluctuated a lot around days 200-300 which corresponds to the end of year 2007 and beginning of 2008, or the so called financial crisis of 2007. At that time, overall market volatility increased as investors lost their trust in the financial institutions. Uncertainty about the risk of the default risk banks contributed greatly to the spike in RV_t. Accordingly, at such turbulent times, higher premiums would have been attached to the asset in question.

Moreover, one can see a surge in RV_t around day 500, which corresponds to the end of 2008 and beginning of year 2009. This can be related to the global pandemic of the so-called swine-flu which, not unlike covid-19, stroke the financial world hard and merciless, causing a decrease in investor confidence and a rise in volatility of stocks. This would also be a sign of destabilisation of the asset [WallStreetMojo].

The character of the RV_t plot reveals also that the shocks seem to echo with a decreasing power thought the time after observation 500 and almost disappear completely until the next shock to the market causes an abrupt increase in volatility again. However, in a build-up to the financial crisis 2007 one can see that volatility shocks explode and gain on magnitude to culminate in a single day where volatility was the highest (ca day 280). Existence of extreme values implies that outliers will have to be taken care of in the context of calculating a linear regression in section 5.

A common thing finance students do is to look for calm and hectic period in volatility on the markets for different assets by eyeballing plots of time series. By looking at Figure 1 created on daily frequency data, it is tempting to try and identify stable and destabilised periods. However, even during the calm periods, it is possible to further divide the observations into arbitrarily calm and hectic semi-periods. My point is that eyeballing is prone to being biased as people in general tend to see patterns where there are none.

Another thing to keep in mind is that realised variance has its limitations when it comes to forecasting possibilities. To name a few :

- RV statistic and every statistic that builds upon it are not forward-looking. Relevant future shocks remain unaccounted for.
- RV calculations are directionless. i.e upward and downward trends in price movements are used in calculations. [WallStreetMojo] Since investors are usually interested in the downside risk, some may take into account only downside price movements. Part of the information would be lost.
- Asset prices are assumed to reflect all available information relevant to measuring volatility, which in turn is defined as standard deviation of the prices over a given time horizon. Usually, more factors then just prices affect volatility.

Having said that, a simple realised variance is is a very precise measure of daily volatility and can be used in heterogeneous autoregressive realised variance (HAR-RV) models.

2 Realised volatility

Realised volatility is just a square root of realised variance [Zamojski, 2020].

With the same provided data, $\sqrt{RV_t}$ was calculated and plotted on daily, 30-minute basis.

As can be seen in Figure 2, the pattern of volatility variation is very similar in magnitude, shape and direction as the one that can be seen in Figure 1. Largest peaks in $\sqrt{RV_t}$ of the prices of the asset are concentrated around the value of 0.04 and occurred during the 2007 financial crisis as well as in 2009, hypothetically, under the swine-flu pandemics.

At the same time, 2 pictures more clearly the build-up leading to the highest fluctuations as the spikes between observations 200-300 seem to perpetuate themselves to a larger extent. In the days following the largest spike (300-500), the fluctuations were varying a lot and does not seem to exhibit any particular pattern.

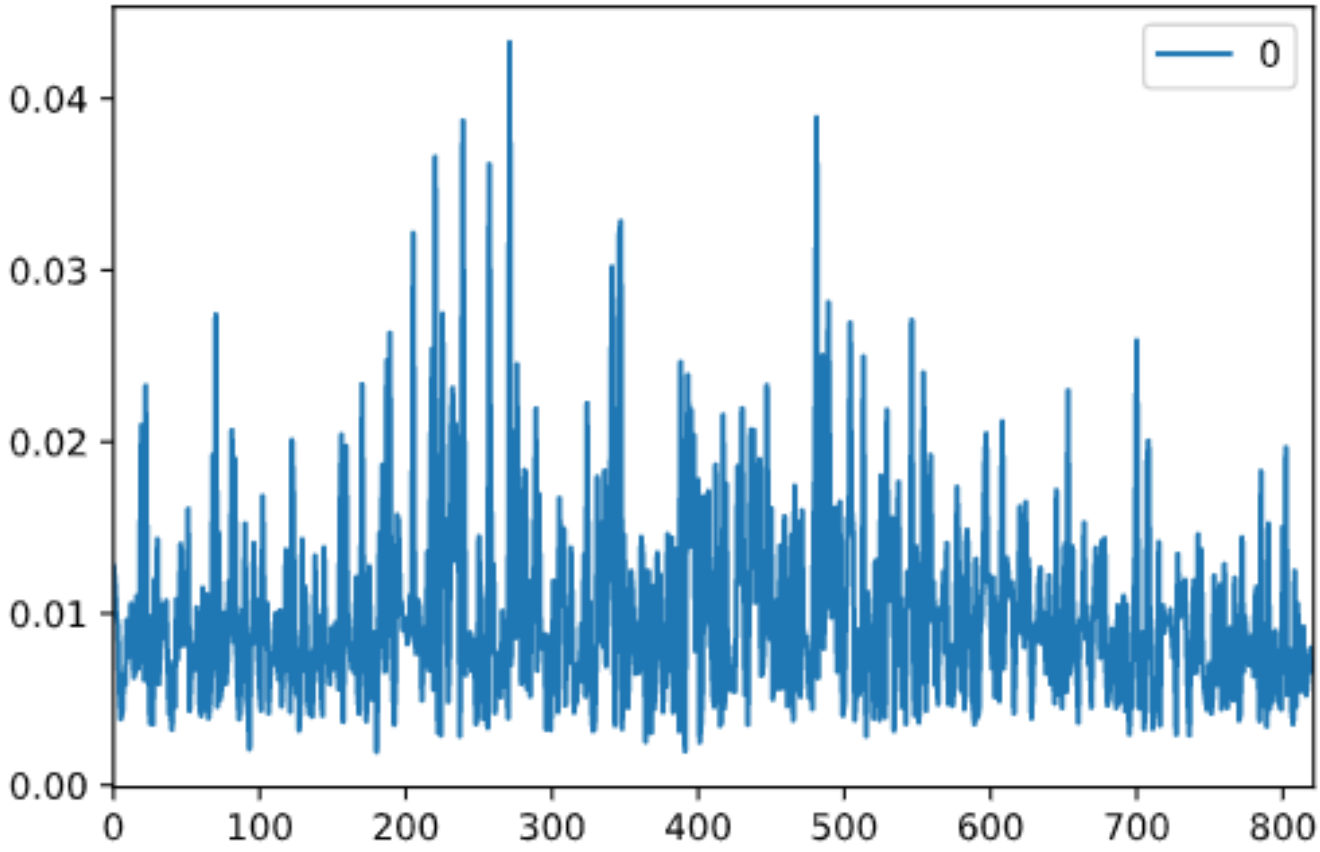

Figure 2: Realised volatility between January 2007 and June 2010.

3 Log realised volatility

Taking logs of the $\sqrt{RV_t}$, we can see that the magnitude, shape and direction of the fluctuations in prices change drastically as compared to the previous two graphs (1,2). No clear pattern can be observed and the many spikes in the volatility that I attributed to financial crisis of 2007 (1, 2) dissappear into the general pattern of fluctuations. Scale used in figure 3 also changes and is a consequence of the logarithmic nature of the transformed prices.

However, the magnitude of changes ranges seem to be larger then the magnitudes of changes in figures 1,2 This can indicate that using log transforms makes the fluctuations larger, and thus, easier to identify.

Furthermore, most of the volatility spikes are extremely short-lived, which may imply that the market in reality re-priced the asset quickly following the shocks. This was hidden in the data in figures 1,2.

The pattern could be seen as a process vaguely similar to some random walk process. Certain random walk models are said to be a special case of the autoregressive models, in which the process reverts to its mean fairly quickly. This notion is developed in section 5, but logarythmic

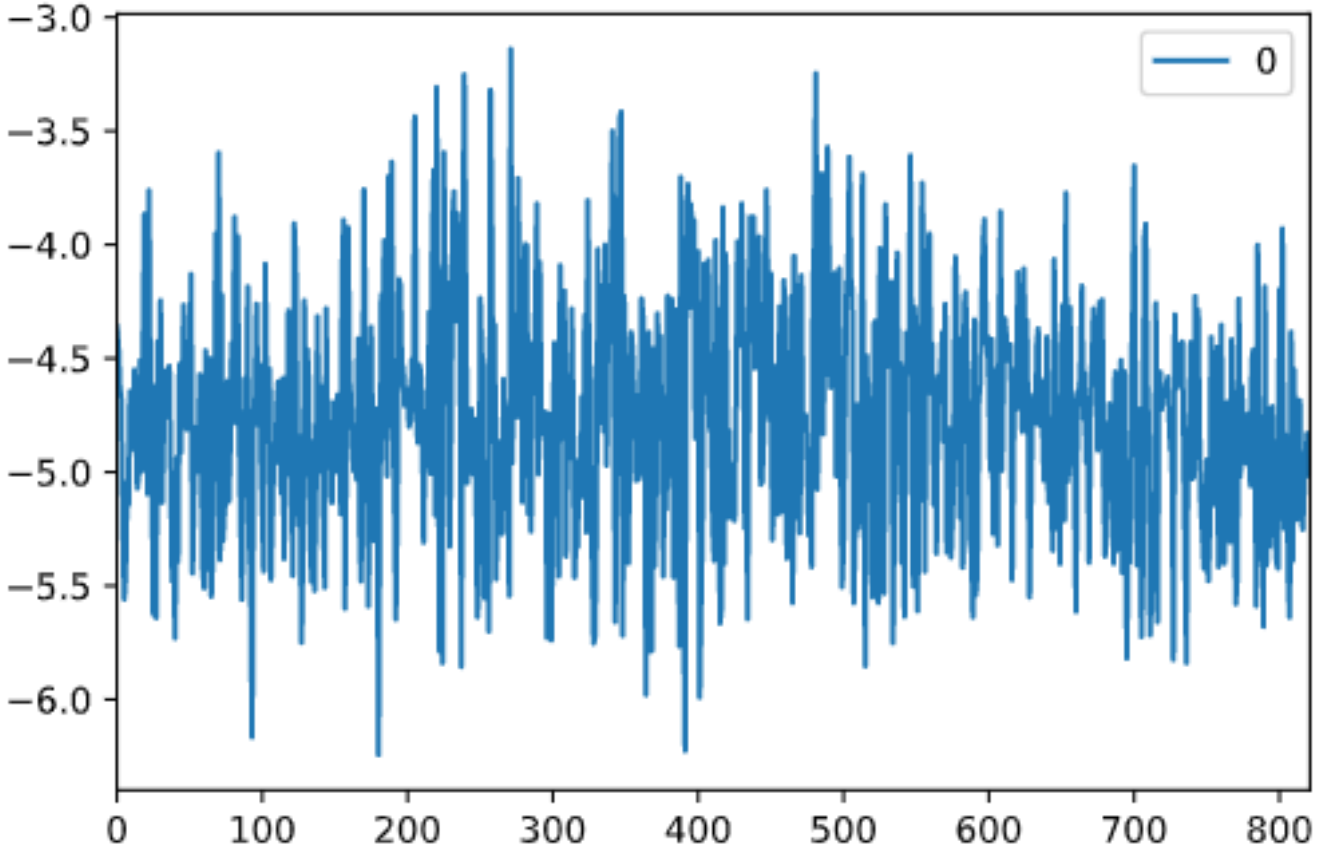

Figure 3: Log of realised volatility between January 2007 and June 2010.

transformation smooths out the time series in question.

The point above suggest that the information we can gather from $ln(\sqrt{RV_t})$ is useful when it comes to forecasting volatility since random walk processes are know to exhibit some predictive power.

4 Comparison

Without going into detail on how to optimise the usage of realised-volatility related methods, it can be said that volatility is time varying : it is persistent and therefore predictable to some extent [Visser, 2011].

Heterogeneous autoregressive realised variance model (HAR-RV) builds upon that idea. In HAR-RV strong persistence typically observed in realised variance of prices can be captured with ordinary least squares (OLS) [Clements and Preve, 2019].

Choice of the estimation scheme, data transformation, and volatility proxy is an issue investigated by several papers [Clements and Preve, 2019].

There is some evidence of that non-linear HAR-models perform better than the raw RV linear

models. According to Tse [2019], the $ln(\sqrt{RV})$ transformation is much closer to a normal distribution then that of pure RV or $\sqrt{RV}$, even if there is still strong evidence against normality (conditional heteroscedasticity in the residuals is reduced but not removed).

Moreover, logarithmic transformation reduces the impact of large observations (extreme deviations from the mean, outliers), which are definitely present in the data. Thus, using log transformed statistics improves the performance of OLS estimators.

To quote Clements and Preve [2019], given that raw RV is prone to spikes/outliers, conditional heteroskedasticity and non-Gaussianity, and the well- known properties of OLS (highly sensitive to outliers, suboptimal in the presence of conditional heteroskedasticity or non-Gaussianity), using realised variance or non-transformed realised volatility is not an optimal choice when alternatives are present. Using $ln(\sqrt{RV})$ in the HAR model is therefore determined to be the most reasonable choice.

5 HAR-RV

To begin with, all three of the above statistics rely on realised variance, so the bias caused by autocorrelation in intraday returns would persist in all of the above regressions 1, 2, 3. This issue has been investigated by, among others, Hansen and Lunde [2006] , who proposed a Newey-West type correction of the RV that yields an unbiased measure of volatility by incorporating the empirical autocovariances into the RV equation. Newly-West standard errors will be thus used in the following analysis.

Models building upon volatilities generated by a certain market component can be described as "autoregressive models reparameterized in a parsimonious way by imposing economically meaningful restrictions." [Corsi, 2009]. One could specify an general autoregressive model AR(p), defined as following :

$$y_i = \delta + \sum_{i=1}^{p} \theta_i y_{t-1} + \varepsilon_t \tag{2}$$

where the current price y_i is equal to a constant δ plus θ times it's previous value, accounting for innovations ε_t [Verbeek, 2008].

For the HAR(3)-RV model, since monthly realized volatility is used (corresponding to 22 working days), the corresponding unrestricted autoregressive model would be an AR(22) [Corsi, 2009].

The linear estimator of $\beta = (b_0, b_1, ..., b_n)$ given the observations $RV_1, ..., RV_n$ is the solution to the minimisation problem [Tse, 2019, Clements and Preve, 2019, Zamojski, 2020]:

$$min_{b_0,b_1,\ldots,b_n} \sum_{t=1}^{n} (RV_t - b_0 - b_1 RV_{t-1}^{daily} - b_2 RV_{t-1}^{weekly} - b_3 RV_{t-1}^{monthly})^2. \quad (3)$$

Assuming that the errors u_t are independent, normally (Gaussian) distributed, and asymptotically homoskedastic, the estimators of β_n are also BLUE estimators.

Following output was obtained:

Dep. Variable:	forecast	**R-squared:**	0.025
Model:	OLS	**Adj. R-squared:**	0.021
Method:	Least Squares	**F-statistic:**	85.18
Date:	Tue, 31 Mar 2020	**Prob (F-statistic):**	8.41e-48
Time:	21:49:08	**Log-Likelihood:**	5772.9
No. Observations:	800	**AIC:**	-1.154e+04
Df Residuals:	796	**BIC:**	-1.152e+04
Df Model:	3		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	6.769e-05	1.07e-05	6.335	0.000	4.68e-05	8.86e-05
daily	0.0569	0.016	3.637	0.000	0.026	0.088
monthly	0.3751	0.079	4.737	0.000	0.220	0.530
weekly	0.0477	0.027	1.740	0.082	-0.006	0.101

Details of the regression can be found in the appendix 5.

Since the data used was transformed into weekly and monthly realised variances, we loose the initial 21 observations as the first five are needed to estimate the first weekly moving average of RV and the rest of them is needed to calculate the first monthly moving average.

The economic interpretation of the model 3 is that it allows economists to obtain insight into the relative importance of past volatilities on the daily volatility.

The table above reports the results of the estimation of the HAR(3)-RV model. Looking at the z-statistics, it can be said that the the realised volatilities aggregated over daily and monthly horizons are very significant, even on $\alpha = 0.01$. The only exception is the coefficient on the weekly realised volatility which would only be significant on $\alpha = 0.1$ level.

It is tricky to find an explanation for the above results. An intuitive approach would suggest that daily estimations should be noisier then weekly estimations and thus of a smaller significance for the model. Instead, the weekly volatility component was estimated to be least significant for forecasting purposes.

At the same time, monthly realised volatilities, being averages over longer periods, should contain less microstructure noise and convey information on the volatility process in a less noisy way. That would explain the magnitude of the coefficient on monthly component.

Guidolin and Pedio [2018] recommends that to evaluate forecasting accuracy, a back-testing exercise can be performed. In such case an HAR model would be obtained using the first half of the intended data, and then a forecast were to be computed with the obtained coefficients for the remaining half of the data. An attempt to do that can be found in the code 5 in the appendix. Due to time limitation and insufficient programming skills this evluation couldn't be successfully performed.

References

A. Clements and D. Preve. A practical guide to harnessing the har volatility model. *Available at SSRN 3369484*, 2019.

F. Corsi. A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):174–196, 2009.

M. Guidolin and M. Pedio. Chapter 10 - realized volatility and covariance. In M. Guidolin and M. Pedio, editors, *Essentials of Time Series for Financial Applications*, pages 381 – 397. Academic Press, 2018. ISBN 978-0-12-813409-2. doi: https://doi.org/10.1016/B978-0-12-813409-2.00010-8. URL `http://www.sciencedirect.com/science/article/pii/B9780128134092000108`.

P. R. Hansen and A. Lunde. Realized variance and market microstructure noise. *Journal of Business & Economic Statistics*, 24(2):127–161, 2006.

Y.-K. Tse. Editorial for the special issue on financia econometrics. *Journal of Risk and Financial Management*, 12(3):153, Sep 2019. ISSN 1911-8074. doi: 10.3390/jrfm12030153. URL `http://dx.doi.org/10.3390/jrfm12030153`.

M. Verbeek. *A guide to modern econometrics*. John Wiley & Sons, 2008.

M. P. Visser. Garch parameter estimation using high-frequency data. *Journal of Financial Econometrics*, 9(1):162–197, 2011.

WallStreetMojo. Realized volatility. `https://www.wallstreetmojo.com/realized-volatility/`. Accessed: 2020-03-28.

M. Zamojski. Lecture notes in financial econometrics, lectures 8-10, February-March 2020.

Appendices

A Regression details

Omnibus:	743.498	**Durbin-Watson:**	1.991
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	25702.435
Skew:	4.230	**Prob(JB):**	0.00
Kurtosis:	29.448	**Cond. No.**	2.49e+04

Warnings:

[1] Standard Errors are heteroscedasticity and autocorrelation robust (HAC) using 1000 lags and without small sample correction

[2] The condition number is large, 2.49e+04. This might indicate that there are strong multicollinearity or other numerical problems.

B Code

```
1
2  import pandas as pd
3  import numpy as np
4  import matplotlib as plt
5  import glob
6  import csv
7
8
9  print(glob.glob('.../data/*.csv'))
10
11 #calculate daily realised variance
12 RVstatsq_1 = []
13
14 for filename in glob.glob('.../data/*.csv'):
15     df = pd.read_csv(filename, parse_dates =["TIMESTAMP"], index_col ="TIMESTAMP")
16     df.resample('30T', label='right', closed='right').last()
17     df['log_price'] = np.log(df['PRICE'])
18     df['daily_returns'] = (df['log_price']- df['log_price'].shift(1))
19     df.dropna(inplace = True)
20     df['returnssq'] = (df['daily_returns']**2)
21     RVstatsq = df['returnssq'].sum()
```

```
    RVstatsq_1.append(RVstatsq)

pd.DataFrame(RVstatsq_1).to_csv("RVstatsq.csv")

#plot the results

df1 = pd.read_csv('.../RVstatsq.csv', header = None)
df1[:10]
df1.plot()


#calculate daily realised volatility

RVstat_2 = []
for filename in glob.glob('.../data/*.csv'):
    df = pd.read_csv(filename, parse_dates =["TIMESTAMP"], index_col ="TIMESTAMP")
    df.resample('30T', label='right', closed='right').last()
    df['log_price'] = np.log(df['PRICE'])
    df['daily_returns'] = (df['log_price']- df['log_price'].shift(1))
    df.dropna(inplace = True)
    df['returnssq'] = (df['daily_returns']**2)
    RVstatsq = df['returnssq'].sum()
    RVstat = RVstatsq**0.5
    RVstat_2.append(RVstat)


print(RVstat_2)
pd.DataFrame(RVstat_2).to_csv("RVstat.csv")

df2 = pd.read_csv('.../RVstat.csv', header = None)
df2[:10]
df2.plot()

#calculate log daily realised variance

logRVstat_1 = []

for filename in glob.glob('.../data/*.csv'):
    df = pd.read_csv(filename, parse_dates =["TIMESTAMP"], index_col ="TIMESTAMP")
    df.resample('30T', label='right', closed='right').sum()
```

```
    df['log_price'] = np.log(df['PRICE'])
    df['daily_returns'] = (df['log_price']- df['log_price'].shift(1))
    df.dropna(inplace = True)
    df['returnssq'] = (df['daily_returns']**2)
    RVstatsq = df['returnssq'].sum()
    RVstat = RVstatsq**0.5
    logRVstat = np.log(RVstat)
    print(logRVstat)
    logRVstat_1.append(logRVstat)

pd.DataFrame(logRVstat_1).to_csv('logRVstat.csv')

df3 = pd.read_csv('.../logRVstat.csv', header = None)
df3[:10]
df3.plot()

#HAR-RV

#transform the dependent variables

df4 = pd.read_csv('.../RVstatsq.csv', header = None, names=['index', 'daily'])
df4[:10]


#weekly RV

df4['weekly'] = df4['daily'].rolling(window=5, min_periods=5).mean()
df4[:30]
df4.dropna(inplace = True)


#monthly RV
df4['monthly'] = df4['daily'].rolling(window=22, min_periods=22).mean()
df4[:30]
df4.dropna(inplace=False)


#OLS

import statsmodels.api as sm
```

```
102
103 #forecasted values
104
105 df4['forecast']=df4['daily'].shift(-1)
106 df4.dropna()
107 df4.head(50)
108 print(df4)
109 df4.dropna(inplace=True)
110
111 #estimate coefficients
112 X = df4[['daily','monthly','weekly']]
113 Y = df4['forecast']
114
115 X = sm.add_constant(X)
116
117 model = sm.OLS(Y, X).fit(cov_type='HAC',cov_kwds={'maxlags':1000})
118
119
120 model_details = model.summary()
121 print(model_details)
122 print(model.summary().as_latex())
123
124
125 #check the forecasting accuracy
126
127 #split dataframe in two parts
128 def split(df, headSize) :
129     hd = df.head(headSize)
130     tl = df.tail(len(df)-headSize)
131     return hd, tl
132
133 first , second = split(df4, 400)
134
135 first.head(100)
136 second.head(100)
137
138 #run HAR on the first part
139
140 X = first[['daily','monthly','weekly']]
141 Y = first['forecast']
```

```
142
143 X = sm.add_constant(X)
144
145 model2 = sm.OLS(Y, X).fit(cov_type='HAC',cov_kwds={'maxlags':1000})
146
147
148 model2_details = model.summary()
149 print(model2_details)
150
151 #estimate a forecast with the obtained coefficients
152
153 nsample = 400
154 sig = 1
155 X = second[['daily','monthly','weekly']]
156 X = sm.add_constant(X)
157 beta = [0.0569, 6.769e-05, 0.3751, 0.04770, ]
158 y_true = np.dot(X, beta)
159 Y = y_true + sig * np.random.normal(size=nsample)
160
161 olsmod = sm.OLS(Y, X)
162 olsres = olsmod.fit()
163 print(olsres.summary())
164
165 #assume standard normal error
166 test = []
167 nsample = 400
168 sig = 1
169 test = 6.769e-05 + 0.0569*first['daily'].shift(-1) + 0.3751*first['monthly'].shift
    (-1) + 0.04770*first['weekly'].shift(-1) + sig * np.random.normal(size=nsample)
170 test.head(100)
171
172 pd.DataFrame(test).to_csv('test.csv')
173 test.head(100)
174
175 #compare
176
177 test['actual'] = pd.concat(first['daily'], axis=1)
178 test.head(400)
```