Introduction to High Frequency Trading Data Analysis

Erphan Al-Delgir, Amber Brodeur, Fengdi (Katrina) Jiao, and Ya Hui (Katie) Zheng

05/01/2017

Contents

Introduction

High frequency trading (HFT) is a type of high-volume trading in multiple markets at high speeds, where small profits are made. Profits are made by capitalizing on price differences in different markets. Trading large volume results in big profits. Data can be recorded in instances of weekly, monthly, daily, hourly, minutely, or fractional-minute intervals. This topic is important because there are controversial claims and risks of HFT. Provides important insight into trading processes and market microstructures. We've only learned about time series methods for monthly and quarterly data. There are new assumptions and parameters to consider as the frequency of data increases. High Frequency trading can be used to comparing efficiency in trading systems, dynamics of bid / ask quotes, market liquidity insights, and assessing realized volatility.

The four largest risks of high frequency trading are:

1. Amplification of systematic risk
2. Errant algorithms
3. Huge investor losses
4. Loss of confidence in market strategy [1]

Controversial claims of high frequency trading are:

- Replaced many broker-deals
- Human interaction and decision making removed
- Mathematical models and algorithms make decisions
- Decisions made in a small fraction of a second
- Large companies profit at the expense of small companies and institutional/retail investors
- Causes "ghost liquidity" [2]

[1] http://www.investopedia.com/articles/markets/012716/four-big-risks-algorithmic-highfrequency-trading.asp

[2] http://www.investopedia.com/articles/markets/012716/four-big-risks-algorithmic-highfrequency-trading.asp

Identifying microstructures: By observing data at finer observations, we are able to see more nuances in autocorrelation, trend, and seasonality that are not detected in in monthly, quarterly, and annual data.

Nonsynchronous observations: At the most finite levels, bid and ask prices on stock exchanges do not happen at fixed intervals.

- Data may either be homogenous or heterogeneously spaced
- Possible Solution: Record observations and serialize to apply time series model using a "regularizing operator" to homogenize intervals.

Objective of the Model

The objective of the high frequency analysis is to analyze high frequency trades in the Dow Jones. More specifically, the daily Dow Jones Realized Variance in 2008 is investigated using the heterogeneous autoregressive model for realized volatility (HARRV). To accomplish the objective, the `highfrequency` package in R software is used. The `highfrequency` package is used to, "Provide functionality to manage, clean and match high frequency trades and quotes data, calculate various liquidity measures, estimate and forecast volatility, and investigate microstructure noise and intraday periodicity."[3] In this tutorial, the heterogenous autoregressive model for realized volatility (HARRV) is explored. The R Documentation describes the HARRV model as being "mainly used to forecast the next days' volatility based on the high-frequency returns of the past." [4]

Model Assumptions

Error term is white noise (zero-mean, serially uncorrelated) → will make the regressors uncorrelated with future error terms zero mean, independent, identically distributed (iid)

The hierarchical structure assumed in the HAR model includes three partial components:

- short-term traders with daily or higher trading frequency
- medium-term traders with weekly trading frequency
- long-term traders with monthly or lower trading frequency
- Therefore, the lag structure in the HAR is fixed as (1, 5, 22) by default

The conditional expectation of tomorrow's realized variance is the weighted sum (plus an intercept) of daily, weekly, and monthly realized volatilities –> So all coefficients B (d) , B (w) , B (m) should be positive.

Modeling with HF Data

Stock price estimation

- bid-ask spreads, transaction pricing, etc.

Return estimation

- relative returns of a security using logarithmic prices

Realized volatility

- determining future expected volatility using historical data with "long memory"

Convolution Operators (for non-homogenous data)

[3] R Documentation; highfrequency-package{highfrequency}

[4] R Documentation; highfrequency-package{highfrequency}

- Exponential Moving Average
- Moving Norm, Variance, and Standard Deviation
- Moving correlation

Dataset Description

The dataset used is called `realized_library` and can be found in the `highfrequency` library in R. The `realized_library` dataset contains data on the daily returns, daily realized variance, and daily realized kernels for different indices and exchange rates from the Oxford-Man Institute of Quantitative Finance. The dataset is an xts object that ranges from 01/03/96 to 03/01/09.

In this analysis, the Dow Jones Realized Variance in 2008 is investigated using the heterogeneous autoregressive model for realized volatility (HARRV).

Model Preparation

The `xts` library stands for "extensible time series" and is used to manage time series data.

The `highfrequency` library is used to, "Provide functionality to manage, clean and match high frequency trades and quotes data, calculate various liquidity measures, estimate and forecast volatility, and investigate microstructure noise and intraday periodicity."[5]

Load the libraries.

```r
library(xts)
library(highfrequency)
```

The line of code below retrieves the sample daily realized volatility data from the `highfrequency` package.

```r
data(realized_library)
```

The functions below are a few basic R functions to learn about the data.

```r
dim(realized_library)
```

```
## [1] 3933   12
```

```r
names(realized_library)
```

```
##  [1] "Dow.Jones.Industrials.Returns"
##  [2] "Dow.Jones.Industrials.Realized.Variance"
##  [3] "Dow.Jones.Industrials.Realized.Kernel"
##  [4] "CAC.40.Returns"
##  [5] "CAC.40.Realized.Variance"
##  [6] "CAC.40.Realized.Kernel"
##  [7] "FTSE.100.Returns"
##  [8] "FTSE.100.Realized.Variance"
##  [9] "FTSE.100.Realized.Kernel"
## [10] "USD.Euro.Returns"
## [11] "USD.Euro.Realized.Variance"
## [12] "USD.Euro.Realized.Kernel"
```

```r
class(realized_library)
```

[5] R Documentation; highfrequency-package{highfrequency}

```
## [1] "xts" "zoo"
```

```
str(realized_library)
```

```
## An 'xts' object on 1996-01-03/2009-03-01 containing:
##   Data: num [1:3933, 1:12] 0.00195 -0.00489 0.00321 0.00439 -0.01139 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : NULL
##   ..$ : chr [1:12] "Dow.Jones.Industrials.Returns" "Dow.Jones.Industrials.Realized.Variance" "Dow.Jo
##   Indexed by objects of class: [POSIXct,POSIXt] TZ: GMT
##   xts Attributes:
##  NULL
```

There are 3,933 observations and 12 variables in the dataframe. Two of the variables are factors and the other two are numeric. The twelve variables are: `Dow.Jones.Industrials.Returns`, `Dow.Jones.Industrials.Realized.Variance`, `Dow.Jones.Industrials.Realized.Kernel`, `CAC.40.Returns`, `CAC.40.Realized.Variance`, `CAC.40.Realized.Kernel`, `FTSE.100.Returns`, `FTSE.100.Realized.Variance`, `FTSE.100.Realized.Kernel`, `USD.Euro.Returns`, `USD.Euro.Realized.Variance`, and `USD.Euro.Realized.Kernel`. The dataset is of class `xts`. The time series begins January 3, 1996 and ends March 1, 2009.

The line of code below selects the Dow Jones Industrials realized variance and returns.

```
DJI_RV = realized_library$Dow.Jones.Industrials.Realized.Variance
DJI_Return = realized_library$Dow.Jones.Industrials.Return
```

Missing values are removed from the realized variance and returns.

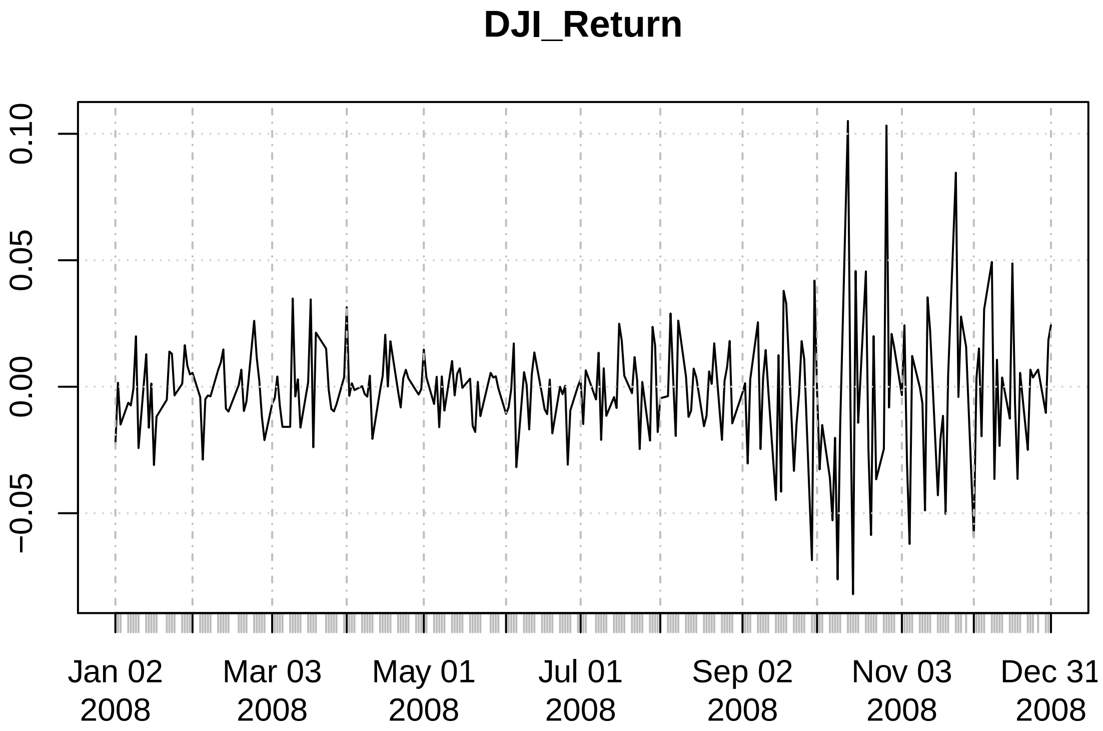
```
DJI_RV = DJI_RV[!is.na(DJI_RV)]
DJI_Return = DJI_Return[!is.na(DJI_Return)]
```

Below, we select the most recent year of data, 2008, from the returns and realized variance

```
DJI_Return = DJI_Return['2008']
DJI_RV = DJI_RV['2008']
```

Below is a plot of the Dow Jones Industrial returns (Daily).

```
plot(DJI_Return)
```



The `aggregatets` function in the `highfrequency` library aggregates a time series object. The command below aggregates `DJI_Return` using the time scale days over five periods, or 5 days. The time period five days is used because five days is one trading week. The result is assigned to `DJI_Return5d`. The `head` command returns the first six observations of `DJI_Return10d`.

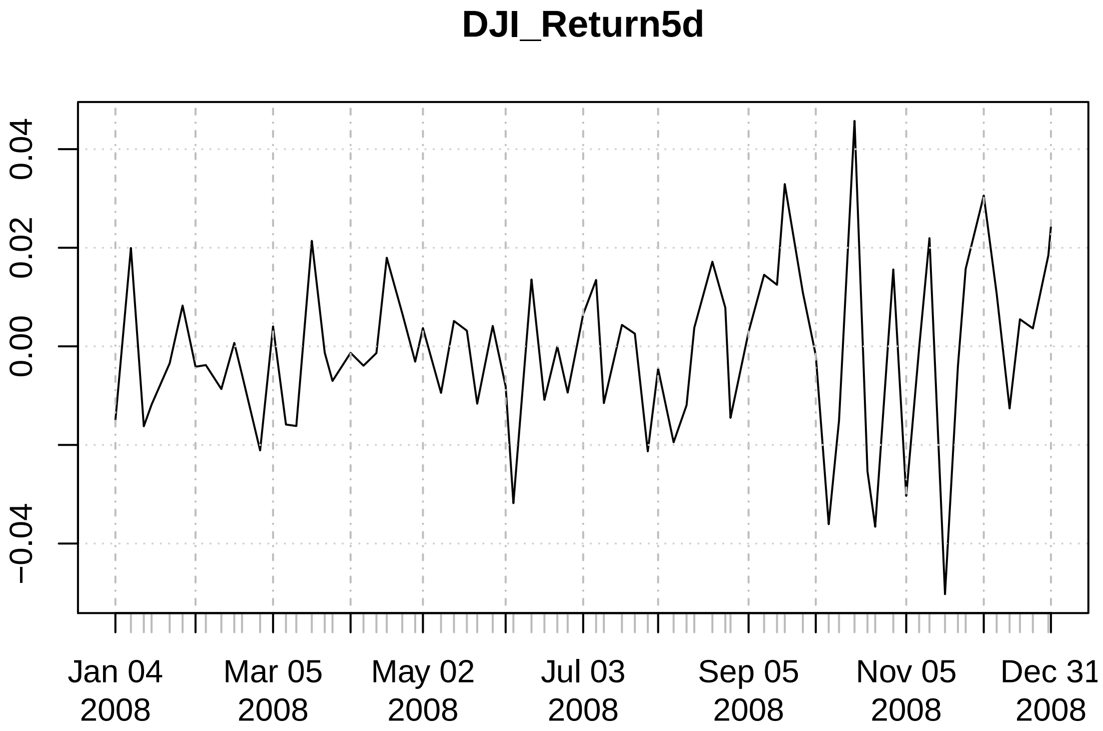
```
DJI_Return5d<- aggregatets(DJI_Return, on="days", k=5,dropna = TRUE)
head(DJI_Return5d)
```

```
## Warning: timezone of object (GMT) is different than current timezone ().
```

```
##            Dow.Jones.Industrials.Returns
## 2008-01-04                  -0.014969042
## 2008-01-10                   0.019952106
## 2008-01-15                  -0.016207472
## 2008-01-18                  -0.011819428
## 2008-01-25                  -0.003432771
## 2008-01-30                   0.008272004
```

Below is a plot of the aggregated Dow Jones Industrial returns using the time scale days over five periods.

```
plot(DJI_Return5d)
```



The summary of DJI_Return5d is below.

```
summary(DJI_Return5d)
```

```
##      Index                        Dow.Jones.Industrials.Returns
##  Min.   :2008-01-04 00:00:00   Min.   :-0.050270
##  1st Qu.:2008-04-05 06:00:00   1st Qu.:-0.011600
##  Median :2008-07-05 12:00:00   Median :-0.001298
##  Mean   :2008-07-04 23:21:04   Mean   :-0.001168
##  3rd Qu.:2008-10-04 18:00:00   3rd Qu.: 0.008160
##  Max.   :2008-12-31 00:00:00   Max.   : 0.045727
```

The rCov function is from the highfrequency package and the functions calculates the realized volatility for the specified return time series at the appropriate frequency. Below, the tick data is aligned to hour to 100 hours.

```
rv5d<- rCov(DJI_Return5d, align.by="hours", align.period=100)
head(rv5d)
```

```
## Warning: timezone of object (GMT) is different than current timezone ().
```

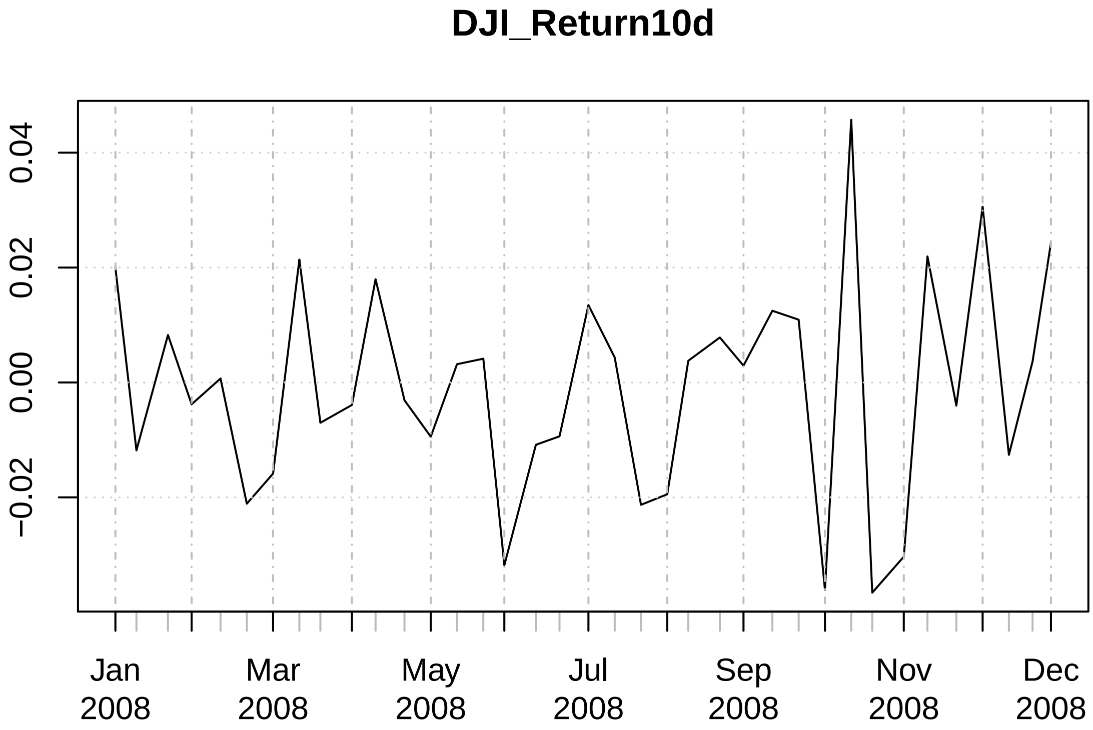
```
##            Dow.Jones.Industrials.Returns
## 2008-01-04                  2.240722e-04
## 2008-01-10                  3.980865e-04
## 2008-01-15                  2.626821e-04
## 2008-01-18                  1.396989e-04
## 2008-01-25                  1.178392e-05
## 2008-01-30                  6.842605e-05
```

The aggregatets function in the highfrequency library aggregates a time series object. The command below aggregates DJI_Return using the time scale days over ten periods, or 10 days. The time period ten days is used because ten days is two trading week. The result is assigned to DJI_Return10d. The head command returns the first six observations of DJI_Return10d.

```
DJI_Return10d<- aggregatets(DJI_Return, on="days", k=10,dropna = TRUE)
```

Below is a plot of the aggregated Dow Jones Industrial returns using the time scale days over ten periods.

```
plot(DJI_Return10d)
```



The summary of `DJI_Return10d` is below.

```
summary(DJI_Return10d)
```

```
##      Index                         Dow.Jones.Industrials.Returns
##  Min.   :2008-01-10 00:00:00   Min.   :-0.0365942
##  1st Qu.:2008-04-09 00:00:00   1st Qu.:-0.0118194
##  Median :2008-07-08 00:00:00   Median : 0.0006900
##  Mean   :2008-07-07 11:01:37   Mean   :-0.0008361
##  3rd Qu.:2008-10-06 00:00:00   3rd Qu.: 0.0109244
##  Max.   :2008-12-31 00:00:00   Max.   : 0.0457272
```

The `rCov` function is from the `highfrequency` package and the functions calculates the realized volatility for the specified return time series at the appropriate frequency. Below, the tick data is aligned to hour to 240 hours.

```
rv10d<- rCov(DJI_Return10d, align.by="hours", align.period=240)
```

The command below aggregates `DJI_Return` using the time scale hours over 528 periods, or 528 hours. The time period 528 hours is used because 528 hours is 22 trading days or 1 trading month. The result is assigned to `DJI_Return22d`. The `head` command returns the first six observations of `DJI_Return22d`.

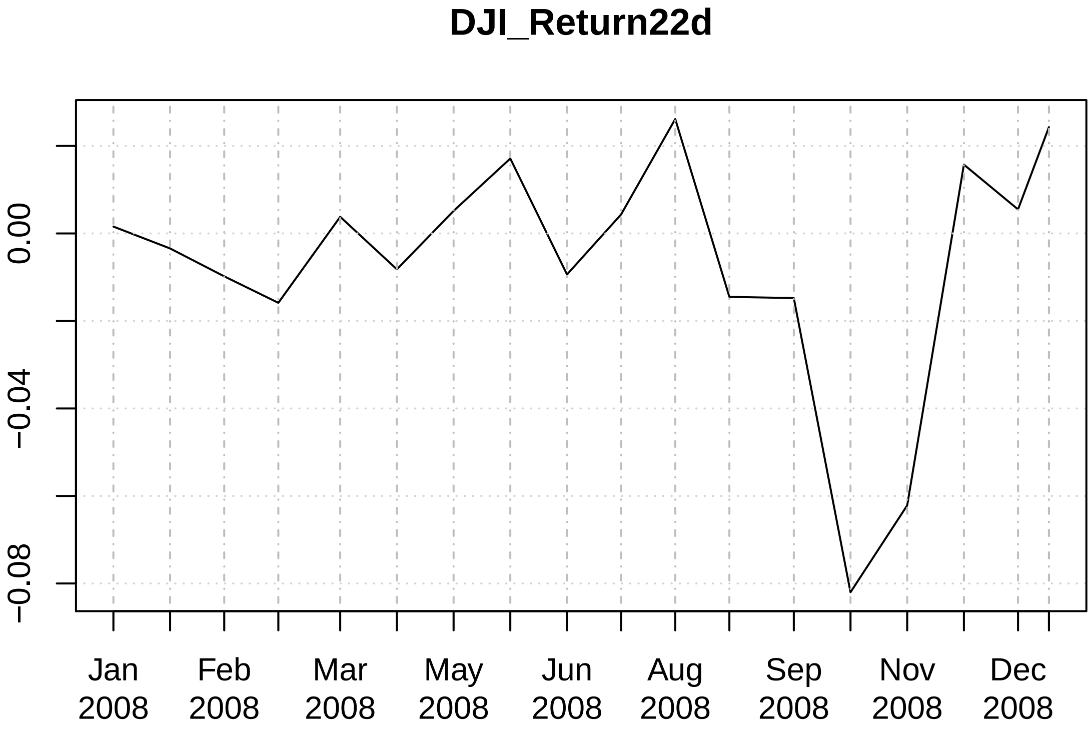
```
DJI_Return22d<- aggregatets(DJI_Return, on="hours", k=528,dropna = TRUE)
head(DJI_Return22d)
```

```
## Warning: timezone of object (GMT) is different than current timezone ().
```

```
##                      Dow.Jones.Industrials.Returns
## 2008-01-03 01:00:00                    0.001567322
## 2008-01-25 01:00:00                  -0.003432771
## 2008-02-15 01:00:00                  -0.009807759
## 2008-03-07 01:00:00                  -0.015858819
## 2008-03-31 01:00:00                   0.003798317
## 2008-04-22 01:00:00                  -0.008204311
```

Below is a plot of the aggregated Dow Jones Industrial returns using the time scale days over ten periods.

```
plot(DJI_Return22d)
```



The summary of `DJI_Return10d` is below.

```
summary(DJI_Return22d)
```

```
##      Index                         Dow.Jones.Industrials.Returns
##  Min.   :2008-01-03 01:00:00   Min.   :-0.0820051
##  1st Qu.:2008-04-05 13:00:00   1st Qu.:-0.0133227
##  Median :2008-07-07 13:00:00   Median :-0.0009327
##  Mean   :2008-07-06 21:00:00   Mean   :-0.0064683
##  3rd Qu.:2008-10-09 13:00:00   3rd Qu.: 0.0054071
##  Max.   :2008-12-31 01:00:00   Max.   : 0.0261513
```

The `rCov` function is from the `highfrequency` package and the functions calculates the realized volatility for the specified return time series at the appropriate frequency. Below, the tick data is aligned to hour to 528 hours.

```
rv22d<- rCov(DJI_Return22d, align.by="hours", align.period=528)
```

Ideally, we would like to calculate returns over short intervals which would allow to increase the sample size of square returns of each day that the realized volatility is calculated. On the other hand, very high-frequency returns introduce microstructure noise that might bias the volatility measure and increase the variability of the volatility measure.

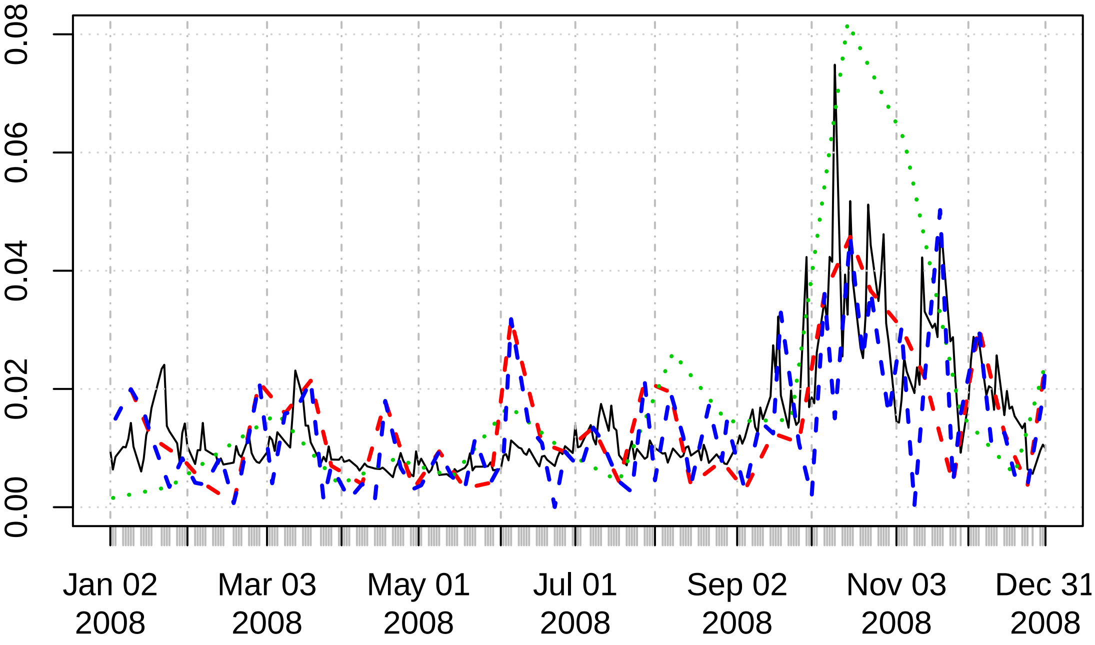
An important choice that has to be made concerns the time interval to use for the returns.

The steps below display the process of determining an appropriate time interval for dataset.

In the example below, we compare realized volatility measures from returns at the 10 seconds, 1 minute, and 10 minutes frequency. Realised Volatility (daily, 5 days ,10 days, and 30 days) : Daily data–Black continuous 5 days–Blue dots 10 days–Red dots 22 days–Green dots

The plot below shows the actual data and the returns for the aggregated 5, 10, and 22 day returns.

```
plot(DJI_RV^0.5,ylim=c(0,0.08),main="", lty=1)  # black continuous
lines(rv10d^0.5,col=2, lty=2, lwd=2)             # red dashed
lines(rv22d^0.5, col=3, lty=3, lwd=2)             # green dots
lines(rv5d^0.5,  col=4, lty=2, lwd=2)             # blue dots
```



The results show that the measure obtained from the 1-day return overall tracks the measures obtained with the higher frequencies, except for some time periods (such as at the end of January, at the beginning of June and at the beginning of October)

- at the end of January: volatility calculated on the 1-day returns spikes up at around 2.5% while the other two measures drop down at around 0.5%
- at the beginning of June: volatility calculated on the 1-day returns spikes around 1% whilst the other two measures stay around 2% on that day.
- at the beginning of October: volatility calculated on the 1-day returns spikes up to almost 8% whilst the other two measures stay below 4.5% on that day.

Overall, there is no optimal way to choose the interval to use. Based on the plot, as shown above, we think that the 5-day frequency looks best to calculate realized volatility measures while the 22-day (monthly) frequency looks worst.

Heterogeneous Autoregressive Model for Realized Volatility (HARRV)

The R Documentation describes the HARRV model as being "mainly used to forecast the next days' volatility based on the high-frequency returns of the past."

The Basic Model:

$RV_{t,t+h} = B_0 \; B_D RV_t + B_w RV_{t\text{-}5,t} + B_M RV_{t\text{-}22,t} + E_{t,t+h}$

- Realized volatility is parameterized as a linear function of the lagged realized volatilities over different horizons
- D: daily; W: weekly; M: monthly
- h: number of days the dependent variable should be aggregated
  - h=1 –> no aggregation takes place → model of daily realized volatility

Model Assumptions:

Error term is white noise (zero-mean, serially uncorrelated) → will make the regressors uncorrelated with future error terms zero mean, independent, identically distributed (iid)

- Three partial components:
  - short-term traders: daily or higher trading frequency
  - medium-term traders: weekly trading frequency
  - long-term traders: monthly or lower trading frequency
  - By default, lag structure in the HAR is fixed as (1, 5, 22)
- All coefficients should be positive
  - Conditional expectation of tomorrow's realized variance is the weighted sum (plus an intercept) of daily, weekly, and monthly realized volatilities

The hierarchical structure assumed in the HAR model includes three partial components:

- short-term traders with daily or higher trading frequency
- medium-term traders with weekly trading frequency
- long-term traders with monthly or lower trading frequency
- Therefore, the lag structure in the HAR is fixed as (1, 5, 22) by default

The conditional expectation of tomorrow's realized variance is the weighted sum (plus an intercept) of daily, weekly, and monthly realized volatilities –> So all coefficients B (d) , B (w) , B (m) should be positive.

The steps below give three models with 1-day, 5-day and 10-day frequency data.

Below is the HARRV model. The `harModel` function is found in the `highfrequency` package, which implements the heterogeneous autoregressive model for realized volatility. The `periods` 1, 5, and 22 are input for one day, week, and month. These inputs are also the default for R. The `RVest` argument is set to `rCor`, or realized volatility. The model type is set to HARRV. The `h` argument sets the dependent variable to be aggregated over one day. The `transform` argument is set to NULL which means no transformation will be performed on the variables.

```
xeg1 = harModel(data=DJI_RV , periods = c(1,5,22), RVest = c("rCov"), type="HARRV",h=1,transform=NULL)
xeg1
```

```
##
## Model:
## RV1 = beta0  +  beta1 * RV1 +  beta2 * RV5 +  beta3 * RV22
##
## Coefficients:
##     beta0      beta1      beta2      beta3
## 4.432e-05  1.586e-01  6.213e-01  8.721e-02
##
##
##     r.squared  adj.r.squared
##        0.4679         0.4608
```

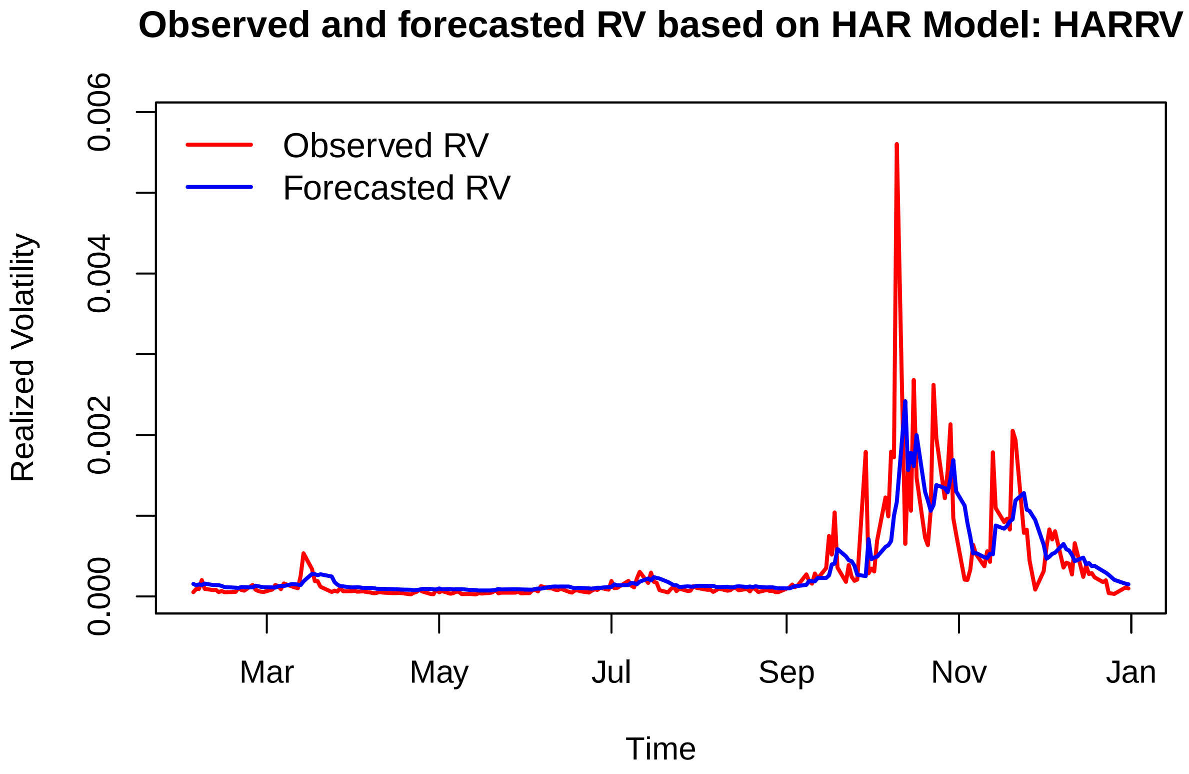
The model summary of `xeg1` is below.

```
summary(xeg1)
```

```
##
## Call:
## "RV1 = beta0  +  beta1 * RV1 +  beta2 * RV5 +  beta3 * RV22"
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0017683 -0.0000626 -0.0000427 -0.0000087  0.0044331
##
## Coefficients:
##        Estimate Std. Error t value Pr(>|t|)
## beta0 4.432e-05  3.695e-05   1.200   0.2315
## beta1 1.586e-01  8.089e-02   1.960   0.0512 .
## beta2 6.213e-01  1.362e-01   4.560 8.36e-06 ***
## beta3 8.721e-02  1.217e-01   0.716   0.4745
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0004344 on 227 degrees of freedom
## Multiple R-squared:  0.4679, Adjusted R-squared:  0.4608
## F-statistic: 66.53 on 3 and 227 DF,  p-value: < 2.2e-16
```

Below is a plot of the HARRV model for `xeg1`.

```
plot(xeg1)
```



The periods 1, 5, and 22 are input for one day, week, and month. These inputs are also the default for R. The `RVest` argument is set to `rCor`, or realized volatility. The model type is set to HARRV. The `h` argument sets the dependent variable to be aggregated over five days. The `transform` argument is set to NULL which means no transformation will be performed on the variables.

```
xeg5=harModel(data=DJI_RV , periods = c(1,5,22), RVest = c("rCov"), type="HARRV",h=5,transform=NULL);
xeg5
```

```
##
## Model:
## RV5 = beta0  +  beta1 * RV1 +  beta2 * RV5 +  beta3 * RV22
##
## Coefficients:
##     beta0      beta1      beta2      beta3
## 7.069e-05  1.955e-01  4.291e-01  1.705e-01
##
##
##     r.squared  adj.r.squared
##        0.5770         0.5713
```

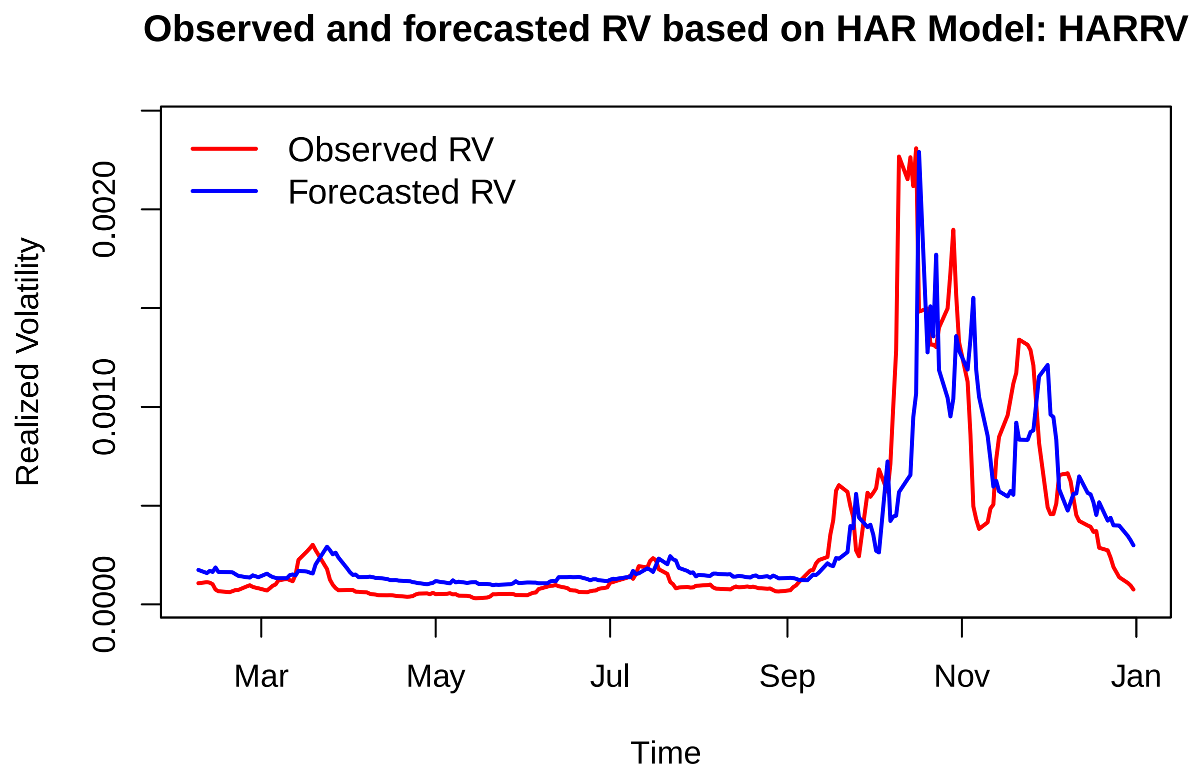
The model summary of xeg5 is below. This is the best time interval for the HARRV model.

```
summary(xeg5)
```

```
##
## Call:
## "RV5 = beta0  +  beta1 * RV1 +  beta2 * RV5 +  beta3 * RV22"
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.055e-03 -8.004e-05 -5.589e-05  4.620e-06  1.699e-03
##
## Coefficients:
##        Estimate Std. Error t value Pr(>|t|)
## beta0 7.069e-05  2.690e-05   2.628 0.009187 **
## beta1 1.955e-01  5.862e-02   3.336 0.000996 ***
## beta2 4.291e-01  9.903e-02   4.333 2.22e-05 ***
## beta3 1.705e-01  8.867e-02   1.923 0.055727 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0003147 on 223 degrees of freedom
## Multiple R-squared:  0.577,  Adjusted R-squared:  0.5713
## F-statistic: 101.4 on 3 and 223 DF,  p-value: < 2.2e-16
```

Below is a plot of the HARRV model for xeg5.

```
plot(xeg5)
```



The periods 1, 5, and 10 are input for one day, one week, and two weeks. These inputs are also the default for R. The `RVest` argument is set to `rCor`, or realized volatility. The model type is set to HARRV. The `h` argument sets the dependent variable to be aggregated over five days. The `transform` argument is set to NULL which means no transformation will be performed on the variables.

```
xeg5.1=harModel(data=DJI_RV , periods = c(1,5,10), RVest = c("rCov"), type="HARRV",h=5,transform=NULL);
xeg5.1
```

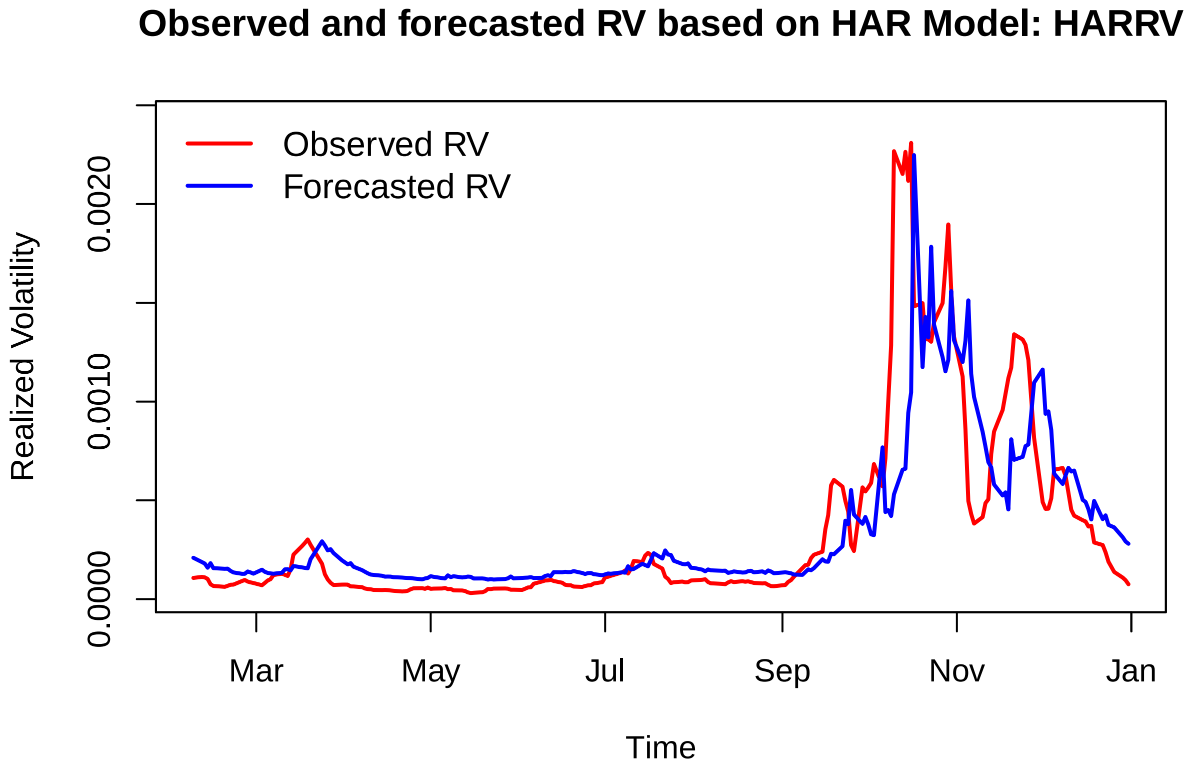
```
##
## Model:
## RV5 = beta0  +  beta1 * RV1 +  beta2 * RV5 +  beta3 * RV10
##
## Coefficients:
##     beta0      beta1      beta2      beta3
## 7.065e-05  2.030e-01  2.142e-01  3.752e-01
##
##
##     r.squared  adj.r.squared
##        0.5855         0.5800
```

```
summary(xeg5.1) #best time interval for the HARRV model
```

```
##
## Call:
## "RV5 = beta0  +  beta1 * RV1 +  beta2 * RV5 +  beta3 * RV10"
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.016e-03 -8.104e-05 -5.442e-05  4.320e-06  1.737e-03
##
## Coefficients:
```

```
##        Estimate Std. Error t value Pr(>|t|)
## beta0 7.065e-05  2.591e-05   2.726 0.006911 **
## beta1 2.030e-01  5.810e-02   3.494 0.000573 ***
## beta2 2.142e-01  1.390e-01   1.541 0.124694
## beta3 3.752e-01  1.295e-01   2.897 0.004142 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0003115 on 223 degrees of freedom
## Multiple R-squared:  0.5855, Adjusted R-squared:   0.58
## F-statistic:   105 on 3 and 223 DF,  p-value: < 2.2e-16
```

```
plot(xeg5.1)
```



The periods 1, 5, and 10 are input for one day, week, and month. These inputs are also the default for R. The RVest argument is set to rCor, or realized volatility. The model type is set to HARRV. The h argument sets the dependent variable to be aggregated over ten days. The transform argument is set to NULL which means no transformation will be performed on the variables.

```
xeg10=harModel(data=DJI_RV , periods = c(1,5,22), RVest = c("rCov"), type="HARRV",h=10,transform=NULL);
xeg10
```

```
##
## Model:
## RV10 = beta0  +  beta1 * RV1 +  beta2 * RV5 +  beta3 * RV22
##
## Coefficients:
##     beta0      beta1      beta2      beta3
## 9.699e-05  1.231e-01  4.410e-01  1.702e-01
##
##
```

```
##     r.squared  adj.r.squared
##        0.5533         0.5472
```

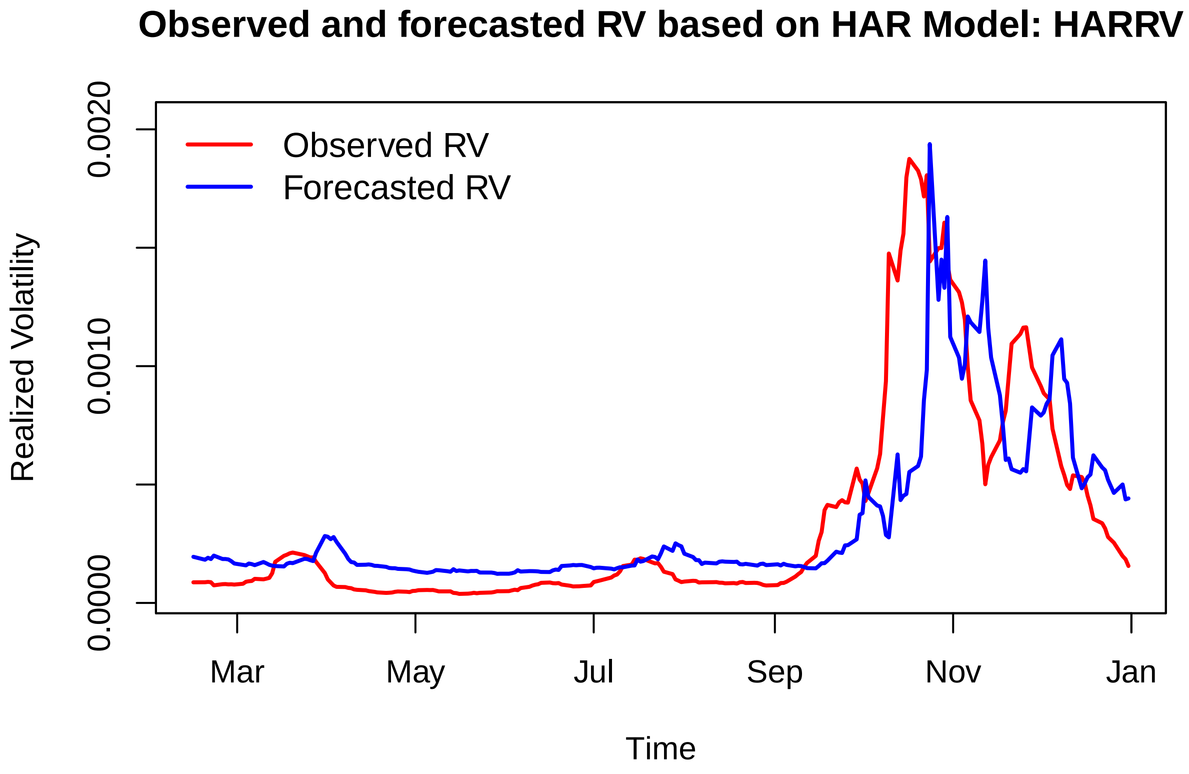
The model summary of `xeg10` is below. This is the best time interval for the HARRV model.

```r
summary(xeg10)
```

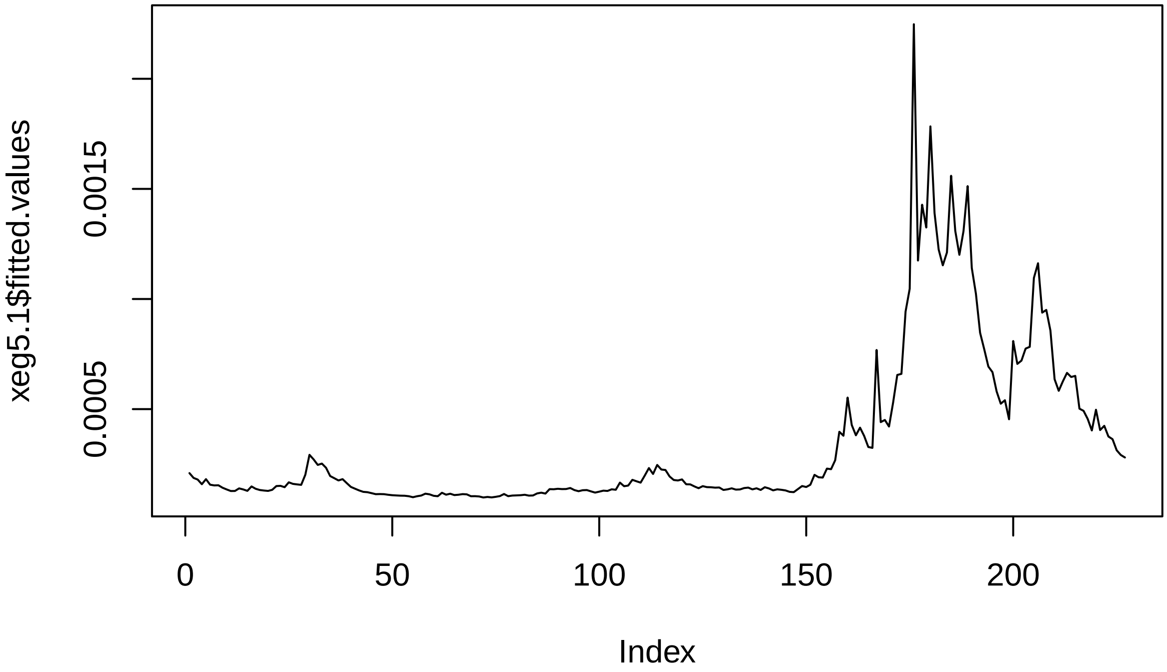
```
##
## Call:
## "RV10 = beta0  +  beta1 * RV1 +  beta2 * RV5 +  beta3 * RV22"
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -9.446e-04 -9.952e-05 -7.484e-05  1.374e-05  1.337e-03
##
## Coefficients:
##        Estimate Std. Error t value Pr(>|t|)
## beta0 9.699e-05  2.607e-05   3.721 0.000253 ***
## beta1 1.231e-01  5.670e-02   2.172 0.030959 *
## beta2 4.410e-01  9.648e-02   4.571 8.14e-06 ***
## beta3 1.702e-01  8.742e-02   1.947 0.052865 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0003043 on 218 degrees of freedom
## Multiple R-squared:  0.5533, Adjusted R-squared:  0.5472
## F-statistic: 90.03 on 3 and 218 DF,  p-value: < 2.2e-16
```

Below is a plot of the HARRV model for `xeg10`.

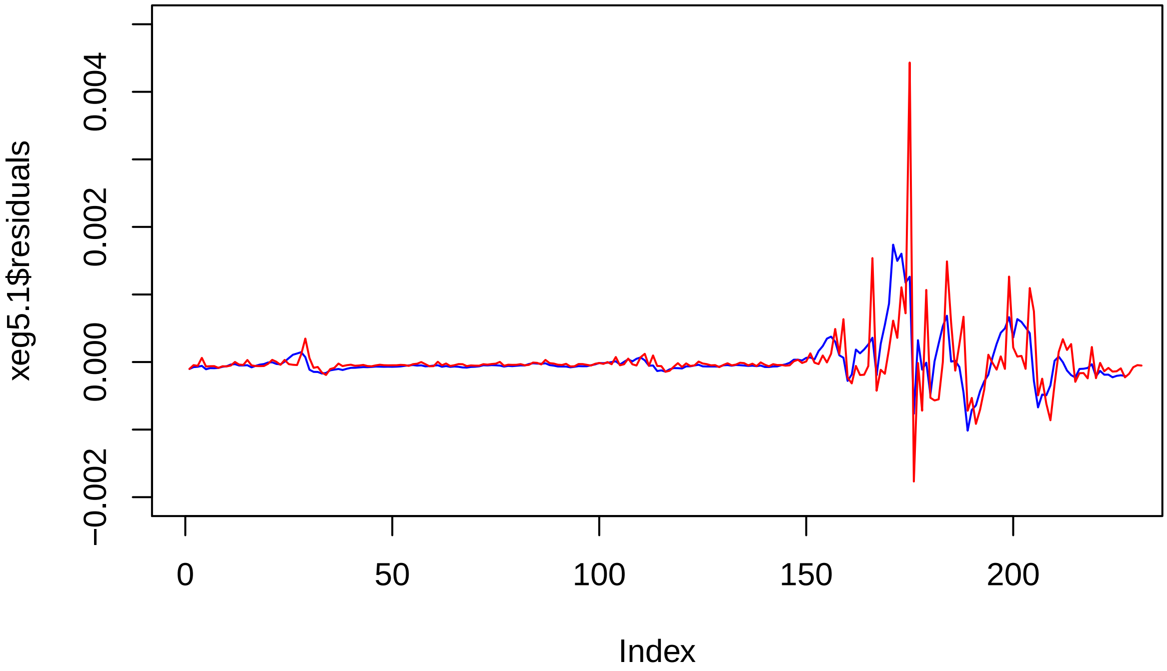
```r
plot(xeg10)
```



```
plot(xeg5.1$fitted.values, type='l')
```



```
plot(xeg5.1$residuals, type='l', col='blue', ylim = c(-.002, 0.005))
lines(xeg1$residuals, type='l', col='red')
```



Conclusion

- HARRV model captured majority of features
- We are left with constant mean and almost constant variance

- Model cannot capture the special events

In the residual plots, we observed a large increase in volatility occuring in November. This model is based on 2008 data, when the world was amidst a financial collapse caused by mortgage backed securities and credit default swaps. At this time, large banks, like AIG and Citigroup, were all going bankrupt and the US government bailed them out to avoid another Great Depression. In essence, the US financial system was bailed out. The increase in volatility in November is most likely due to the huge election year for 2008. The Great Recession was well underway and the world was in search of a hopeful leader.

Works Cited:

1. http://www.investopedia.com/terms/h/high-frequency-trading.asp
2. http://faculty.baruch.cuny.edu/smanzan/FINMETRICS/_book/high-frequency-data.html
3. https://r-forge.r-project.org/scm/viewvc.php/*checkout*/pkg/highfrequency/inst/doc/highfrequency.pdf?revision=37&root=highfrequency&pathrev=38
4. https://www.alexandria.unisg.ch/248495/1/Aud_Hua_Okh_Flexible%20HAR%20Model%20for%20Realized%20Volatility.pdf
5. https://www.researchgate.net/profile/Francesco_Audrino/publication/259146365_Lassoing_the_HAR_Model_A_Model_Selection_Perspective_on_Realized_Volatility_Dynamics/links/0c96052a08f82e14c5000000.pdf
6. https://en.wikipedia.org/wiki/Realized_variance
7. R Documentation; highfrequency-package{highfrequency}
8. http://www.investopedia.com/articles/markets/012716/four-big-risks-algorithmic-highfrequency-trading.asp
9. http://www.investopedia.com/terms/h/high-frequency-trading.asp