STAT430_HW3

Taiga Hasegawa(taigah2)

2019/2/19

Question1

```
weights_fracDiff=function(d,nWei,tau){
  weights=rep(NA,nWei)
  weights[1]=1
  for(i in 2:nWei){
    weight=-weights[i-1]*(d-(i-1)+1)/(i-1)
    if(length(tau)!=0){
      if(abs(weight)>tau){weights[i]=weight}
      else{break}
    }else{
      weights[i]=weight
    }
  }
  return(weights)
}
fracDiff=function(x,d=0.3,nWei=40,tau=NULL){
    weig=weights_fracDiff(d=d,nWei=nWei,tau=tau)
    nWei=length(weig)
    nx=length(x)
    rst=rep(NA,nx)
    rst[nWei:nx]=sapply(nWei:nx,function(i){sum(weig*x[i:(i-nWei+1)])})
    return(rst)
}
```

```
dat <- read.csv("unit_bar_XBTUSD_all.csv", header = T)
dat$V <- as.numeric(dat$V)
dat$C <- as.numeric(dat$C)
trainDat <- dat[1:floor(nrow(dat)/3*2),]
testDat <- dat[(floor(nrow(dat)/3*2)+1):nrow(dat),]
```

```
C_fracD <- fracDiff(x=dat$C,d=0.5,tau=0.001)
idx=!is.na(C_fracD)
cor(dat$C[idx],C_fracD[idx])
```

```
## [1] 0.8256172
```

```
cor(c(0,diff(dat$C))[idx],dat$C[idx])
```

```
## [1] 0.01448753
```

The correlation coefficients for the closed prices of the derived series and the original series were biger than the one for the first order difference of the closed prices and the original series because the d was defined so that memory was preserved.

Question2

```r
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':
##
##     date
```

```r
library(CADFtest)
```

```
## Loading required package: dynlm

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich

## Loading required package: tseries

## Loading required package: urca
```

```r
library(tseries)
library(fUnitRoots)
```

```
## Loading required package: timeDate

## Loading required package: timeSeries

##
## Attaching package: 'timeSeries'

## The following object is masked from 'package:zoo':
##
##     time<-

## Loading required package: fBasics

##
## Attaching package: 'fUnitRoots'

## The following objects are masked from 'package:urca':
##
##     punitroot, qunitroot, unitrootTable
```

```r
library(strucchange)
```

```r
for (i in seq(0.5,0.8,0.1)){
  C_fracD=fracDiff(x=trainDat$C,d=i,tau=0.0001)
  C_fracD=C_fracD[!is.na(C_fracD)]
  print("-------------------------")
  print(i)
  print("-------------------------")
  print(tseries::adf.test(C_fracD))
```

```r
  print(CADFtest::CADFtest(C_fracD, type="trend", max.lag.y=5))
  print(summary(urca::ur.df(C_fracD, type="trend", lags=5)))
  print(tseries::pp.test(C_fracD))
  print(tseries::kpss.test(C_fracD, null="Trend"))
  print("-------------------------")
}
```

```
## [1] "-------------------------"
## [1] 0.5
## [1] "-------------------------"
##
##  Augmented Dickey-Fuller Test
##
## data:  C_fracD
## Dickey-Fuller = -7.3006, Lag order = 14, p-value = 0.01
## alternative hypothesis: stationary
##
##
##  ADF test
##
## data:  C_fracD
## ADF(5) = -10.148, p-value < 2.2e-16
## alternative hypothesis: true delta is less than 0
## sample estimates:
##      delta
## -0.1413239
##
##
## ###############################################
## # Augmented Dickey-Fuller Test Unit Root Test #
## ###############################################
##
## Test regression trend
##
##
## Call:
## lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -744.34  -52.06    2.79   55.52  800.15
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 140.357899  14.710775   9.541  < 2e-16 ***
## z.lag.1      -0.141324   0.013927 -10.148  < 2e-16 ***
## tt           -0.013928   0.002805  -4.965 7.22e-07 ***
## z.diff.lag1  -0.407878   0.020058 -20.335  < 2e-16 ***
## z.diff.lag2  -0.219806   0.020799 -10.568  < 2e-16 ***
## z.diff.lag3  -0.166274   0.020458  -8.128 6.14e-16 ***
## z.diff.lag4  -0.120830   0.019685  -6.138 9.35e-10 ***
## z.diff.lag5  -0.089901   0.017353  -5.181 2.34e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 131.1 on 3276 degrees of freedom
## Multiple R-squared:  0.242,  Adjusted R-squared:  0.2403
## F-statistic: 149.4 on 7 and 3276 DF,  p-value: < 2.2e-16
##
##
## Value of test-statistic is: -10.1477 34.3619 51.5187
##
## Critical values for test statistics:
##       1pct  5pct 10pct
## tau3 -3.96 -3.41 -3.12
## phi2  6.09  4.68  4.03
## phi3  8.27  6.25  5.34
##
##
##  Phillips-Perron Unit Root Test
##
## data:  C_fracD
## Dickey-Fuller Z(alpha) = -1122.5, Truncation lag parameter = 9,
## p-value = 0.01
## alternative hypothesis: stationary
##
##
##  KPSS Test for Trend Stationarity
##
## data:  C_fracD
## KPSS Trend = 1.2943, Truncation lag parameter = 9, p-value = 0.01
##
## [1] "-------------------------"
## [1] "-------------------------"
## [1] 0.6
## [1] "-------------------------"
##
##  Augmented Dickey-Fuller Test
##
## data:  C_fracD
## Dickey-Fuller = -9.0463, Lag order = 14, p-value = 0.01
## alternative hypothesis: stationary
##
##
##  ADF test
##
## data:  C_fracD
## ADF(5) = -13.333, p-value < 2.2e-16
## alternative hypothesis: true delta is less than 0
## sample estimates:
##      delta
## -0.2666339
##
##
## ###############################################
## # Augmented Dickey-Fuller Test Unit Root Test #
## ###############################################
##
```

```
## Test regression trend
##
##
## Call:
## lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -756.37  -51.94    2.96   56.25  787.16
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 145.799584  11.926720  12.225  < 2e-16 ***
## z.lag.1      -0.266634   0.019999 -13.333  < 2e-16 ***
## tt           -0.014229   0.002648  -5.375 8.21e-08 ***
## z.diff.lag1  -0.386559   0.022947 -16.845  < 2e-16 ***
## z.diff.lag2  -0.211715   0.023005  -9.203  < 2e-16 ***
## z.diff.lag3  -0.156234   0.022105  -7.068 1.92e-12 ***
## z.diff.lag4  -0.111049   0.020658  -5.376 8.17e-08 ***
## z.diff.lag5  -0.081482   0.017369  -4.691 2.83e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 130.8 on 3276 degrees of freedom
## Multiple R-squared:  0.3068, Adjusted R-squared:  0.3053
## F-statistic: 207.1 on 7 and 3276 DF,  p-value: < 2.2e-16
##
##
## Value of test-statistic is: -13.3326 59.2705 88.8953
##
## Critical values for test statistics:
##       1pct  5pct 10pct
## tau3 -3.96 -3.41 -3.12
## phi2  6.09  4.68  4.03
## phi3  8.27  6.25  5.34
##
##
##  Phillips-Perron Unit Root Test
##
## data:  C_fracD
## Dickey-Fuller Z(alpha) = -2218.6, Truncation lag parameter = 9,
## p-value = 0.01
## alternative hypothesis: stationary
##
##
##  KPSS Test for Trend Stationarity
##
## data:  C_fracD
## KPSS Trend = 0.82993, Truncation lag parameter = 9, p-value = 0.01
##
## [1] "-------------------------"
## [1] "-------------------------"
## [1] 0.7
## [1] "-------------------------"
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  C_fracD
## Dickey-Fuller = -10.96, Lag order = 14, p-value = 0.01
## alternative hypothesis: stationary
##
##
##  ADF test
##
## data:  C_fracD
## ADF(5) = -16.635, p-value < 2.2e-16
## alternative hypothesis: true delta is less than 0
## sample estimates:
##      delta
## -0.4449295
##
##
## ###############################################
## # Augmented Dickey-Fuller Test Unit Root Test #
## ###############################################
##
## Test regression trend
##
##
## Call:
## lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -775.48  -51.63    3.23   55.29  771.46
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 123.229417   8.736785  14.105  < 2e-16 ***
## z.lag.1      -0.444930   0.026747 -16.635  < 2e-16 ***
## tt           -0.011516   0.002503  -4.601 4.37e-06 ***
## z.diff.lag1  -0.316407   0.027003 -11.717  < 2e-16 ***
## z.diff.lag2  -0.168934   0.025971  -6.505 8.97e-11 ***
## z.diff.lag3  -0.122117   0.024198  -5.046 4.75e-07 ***
## z.diff.lag4  -0.085535   0.021811  -3.922 8.98e-05 ***
## z.diff.lag5  -0.065473   0.017392  -3.765  0.00017 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 130.4 on 3276 degrees of freedom
## Multiple R-squared:  0.3718, Adjusted R-squared:  0.3704
## F-statistic:   277 on 7 and 3276 DF,  p-value: < 2.2e-16
##
##
## Value of test-statistic is: -16.635 92.2498 138.3702
##
## Critical values for test statistics:
##       1pct  5pct 10pct
```

```
## tau3 -3.96 -3.41 -3.12
## phi2  6.09  4.68  4.03
## phi3  8.27  6.25  5.34
##
##
##  Phillips-Perron Unit Root Test
##
## data:  C_fracD
## Dickey-Fuller Z(alpha) = -3006.6, Truncation lag parameter = 9,
## p-value = 0.01
## alternative hypothesis: stationary
##
##
##  KPSS Test for Trend Stationarity
##
## data:  C_fracD
## KPSS Trend = 0.40331, Truncation lag parameter = 9, p-value = 0.01
##
## [1] "-------------------------"
## [1] "-------------------------"
## [1] 0.8
## [1] "-------------------------"
##
##  Augmented Dickey-Fuller Test
##
## data:  C_fracD
## Dickey-Fuller = -12.711, Lag order = 14, p-value = 0.01
## alternative hypothesis: stationary
##
##
##  ADF test
##
## data:  C_fracD
## ADF(5) = -19.574, p-value < 2.2e-16
## alternative hypothesis: true delta is less than 0
## sample estimates:
##      delta
## -0.6530819
##
##
## ###############################################
## # Augmented Dickey-Fuller Test Unit Root Test #
## ###############################################
##
## Test regression trend
##
##
## Call:
## lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -791.51  -50.25    3.00   54.36  762.21
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 79.019966   6.098973  12.956  < 2e-16 ***
## z.lag.1     -0.653082   0.033364 -19.574  < 2e-16 ***
## tt          -0.006619   0.002418  -2.737 0.006232 **
## z.diff.lag1 -0.214991   0.031609  -6.802 1.22e-11 ***
## z.diff.lag2 -0.105027   0.029352  -3.578 0.000351 ***
## z.diff.lag3 -0.075159   0.026529  -2.833 0.004637 **
## z.diff.lag4 -0.053313   0.023039  -2.314 0.020728 *
## z.diff.lag5 -0.047598   0.017412  -2.734 0.006298 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 130 on 3276 degrees of freedom
## Multiple R-squared:  0.4325, Adjusted R-squared:  0.4313
## F-statistic: 356.6 on 7 and 3276 DF,  p-value: < 2.2e-16
##
##
## Value of test-statistic is: -19.5744 127.7234 191.5833
##
## Critical values for test statistics:
##       1pct  5pct 10pct
## tau3 -3.96 -3.41 -3.12
## phi2  6.09  4.68  4.03
## phi3  8.27  6.25  5.34
##
##
##  Phillips-Perron Unit Root Test
##
## data:  C_fracD
## Dickey-Fuller Z(alpha) = -3316, Truncation lag parameter = 9,
## p-value = 0.01
## alternative hypothesis: stationary
##
##
##  KPSS Test for Trend Stationarity
##
## data:  C_fracD
## KPSS Trend = 0.13888, Truncation lag parameter = 9, p-value =
## 0.06319
##
## [1] "--------------------------"
```

When d=0.8, fractionally differentiated closed price passed all the unit root tests and the stationarity test. I also checked whether d=0.79 passed those tests.

```
C_fracD=fracDiff(x=trainDat$C,d=0.79,tau=0.0001)
C_fracD=C_fracD[!is.na(C_fracD)]
tseries::adf.test(C_fracD)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  C_fracD
## Dickey-Fuller = -12.551, Lag order = 14, p-value = 0.01
```

```
## alternative hypothesis: stationary
```

```r
tseries::kpss.test(C_fracD, null="Trend")
```

```
##
##  KPSS Test for Trend Stationarity
##
## data:  C_fracD
## KPSS Trend = 0.15664, Truncation lag parameter = 9, p-value =
## 0.04113
```

In this case, it didn't pass the KPSS test. Therfore, d=0.8 is the appropriate value.

Next we are talking about fractionally differentiated volumes.

```r
for (i in seq(0.5,0.8,0.1)){
  V_fracD=fracDiff(x=trainDat$V,d=i,tau=0.0001)
  V_fracD=V_fracD[!is.na(V_fracD)]
  print("-------------------------")
  print(i)
  print("-------------------------")
  print(tseries::adf.test(V_fracD))
  print(CADFtest::CADFtest(V_fracD, type="trend", max.lag.y=5))
  print(summary(urca::ur.df(V_fracD, type="trend", lags=5)))
  print(tseries::pp.test(V_fracD))
  print(tseries::kpss.test(V_fracD, null="Trend"))
  print("-------------------------")
}
```

```
## [1] "-------------------------"
## [1] 0.5
## [1] "-------------------------"
##
##  Augmented Dickey-Fuller Test
##
## data:  V_fracD
## Dickey-Fuller = -6.874, Lag order = 14, p-value = 0.01
## alternative hypothesis: stationary
##
##
##  ADF test
##
## data:  V_fracD
## ADF(5) = -9.758, p-value < 2.2e-16
## alternative hypothesis: true delta is less than 0
## sample estimates:
##      delta
## -0.1118542
##
##
## ###############################################
## # Augmented Dickey-Fuller Test Unit Root Test #
## ###############################################
##
## Test regression trend
##
```

```
##
## Call:
## lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -13279272   -570107    -35938    544245   8537058
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.359e+05  1.038e+05   9.015  < 2e-16 ***
## z.lag.1     -1.119e-01  1.146e-02  -9.758  < 2e-16 ***
## tt           1.141e+02  2.509e+01   4.549 5.59e-06 ***
## z.diff.lag1 -1.946e-01  1.885e-02 -10.324  < 2e-16 ***
## z.diff.lag2 -1.993e-01  1.880e-02 -10.605  < 2e-16 ***
## z.diff.lag3 -9.479e-02  1.877e-02  -5.049 4.69e-07 ***
## z.diff.lag4 -1.010e-01  1.805e-02  -5.596 2.38e-08 ***
## z.diff.lag5 -1.228e-01  1.733e-02  -7.083 1.71e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1196000 on 3276 degrees of freedom
## Multiple R-squared:  0.1412, Adjusted R-squared:  0.1394
## F-statistic: 76.97 on 7 and 3276 DF,  p-value: < 2.2e-16
##
##
## Value of test-statistic is: -9.758 31.7563 47.6199
##
## Critical values for test statistics:
##       1pct  5pct 10pct
## tau3 -3.96 -3.41 -3.12
## phi2  6.09  4.68  4.03
## phi3  8.27  6.25  5.34
##
##
##  Phillips-Perron Unit Root Test
##
## data:  V_fracD
## Dickey-Fuller Z(alpha) = -579.72, Truncation lag parameter = 9,
## p-value = 0.01
## alternative hypothesis: stationary
##
##
##  KPSS Test for Trend Stationarity
##
## data:  V_fracD
## KPSS Trend = 1.1717, Truncation lag parameter = 9, p-value = 0.01
##
## [1] "-------------------------"
## [1] "-------------------------"
## [1] 0.6
## [1] "-------------------------"
##
##  Augmented Dickey-Fuller Test
```

```
##
## data:  V_fracD
## Dickey-Fuller = -8.5092, Lag order = 14, p-value = 0.01
## alternative hypothesis: stationary
##
##
##  ADF test
##
## data:  V_fracD
## ADF(5) = -12.66, p-value < 2.2e-16
## alternative hypothesis: true delta is less than 0
## sample estimates:
##      delta
## -0.2065853
##
##
## ###############################################
## # Augmented Dickey-Fuller Test Unit Root Test #
## ###############################################
##
## Test regression trend
##
##
## Call:
## lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -13032001   -576777    -42661    532410   8749459
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.626e+05  8.633e+04  11.150  < 2e-16 ***
## z.lag.1     -2.066e-01  1.632e-02 -12.660  < 2e-16 ***
## tt           1.154e+02  2.387e+01   4.833 1.40e-06 ***
## z.diff.lag1 -2.030e-01  2.065e-02  -9.830  < 2e-16 ***
## z.diff.lag2 -1.953e-01  2.028e-02  -9.631  < 2e-16 ***
## z.diff.lag3 -9.382e-02  1.989e-02  -4.718 2.49e-06 ***
## z.diff.lag4 -9.470e-02  1.868e-02  -5.069 4.22e-07 ***
## z.diff.lag5 -1.181e-01  1.734e-02  -6.808 1.17e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1194000 on 3276 degrees of freedom
## Multiple R-squared:  0.1971, Adjusted R-squared:  0.1954
## F-statistic: 114.9 on 7 and 3276 DF,  p-value: < 2.2e-16
##
##
## Value of test-statistic is: -12.6603 53.4355 80.1472
##
## Critical values for test statistics:
##       1pct  5pct 10pct
## tau3 -3.96 -3.41 -3.12
## phi2  6.09  4.68  4.03
```

```
## phi3  8.27  6.25  5.34
##
##
##  Phillips-Perron Unit Root Test
##
## data:  V_fracD
## Dickey-Fuller Z(alpha) = -1197.9, Truncation lag parameter = 9,
## p-value = 0.01
## alternative hypothesis: stationary
##
##
##  KPSS Test for Trend Stationarity
##
## data:  V_fracD
## KPSS Trend = 0.76873, Truncation lag parameter = 9, p-value = 0.01
##
## [1] "-------------------------"
## [1] "-------------------------"
## [1] 0.7
## [1] "-------------------------"
##
##  Augmented Dickey-Fuller Test
##
## data:  V_fracD
## Dickey-Fuller = -10.242, Lag order = 14, p-value = 0.01
## alternative hypothesis: stationary
##
##
##  ADF test
##
## data:  V_fracD
## ADF(5) = -15.643, p-value < 2.2e-16
## alternative hypothesis: true delta is less than 0
## sample estimates:
##      delta
## -0.3404567
##
##
## ###############################################
## # Augmented Dickey-Fuller Test Unit Root Test #
## ###############################################
##
## Test regression trend
##
##
## Call:
## lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -12835843   -567319    -38109    531783   8887915
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  8.258e+05  6.704e+04  12.319  < 2e-16 ***
## z.lag.1     -3.405e-01  2.176e-02 -15.643  < 2e-16 ***
## tt           9.466e+01  2.277e+01   4.158 3.30e-05 ***
## z.diff.lag1 -1.753e-01  2.336e-02  -7.505 7.89e-14 ***
## z.diff.lag2 -1.676e-01  2.240e-02  -7.484 9.23e-14 ***
## z.diff.lag3 -7.556e-02  2.139e-02  -3.532 0.000418 ***
## z.diff.lag4 -7.701e-02  1.951e-02  -3.946 8.10e-05 ***
## z.diff.lag5 -1.065e-01  1.737e-02  -6.133 9.67e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1191000 on 3276 degrees of freedom
## Multiple R-squared:  0.2568, Adjusted R-squared:  0.2552
## F-statistic: 161.7 on 7 and 3276 DF,  p-value: < 2.2e-16
##
##
## Value of test-statistic is: -15.6433 81.5743 122.359
##
## Critical values for test statistics:
##       1pct  5pct 10pct
## tau3 -3.96 -3.41 -3.12
## phi2  6.09  4.68  4.03
## phi3  8.27  6.25  5.34
##
##
##  Phillips-Perron Unit Root Test
##
## data:  V_fracD
## Dickey-Fuller Z(alpha) = -1818.8, Truncation lag parameter = 9,
## p-value = 0.01
## alternative hypothesis: stationary
##
##
##  KPSS Test for Trend Stationarity
##
## data:  V_fracD
## KPSS Trend = 0.39818, Truncation lag parameter = 9, p-value = 0.01
##
## [1] "-------------------------"
## [1] "-------------------------"
## [1] 0.8
## [1] "-------------------------"
##
##  Augmented Dickey-Fuller Test
##
## data:  V_fracD
## Dickey-Fuller = -11.811, Lag order = 14, p-value = 0.01
## alternative hypothesis: stationary
##
##
##  ADF test
##
## data:  V_fracD
## ADF(5) = -18.328, p-value < 2.2e-16
```

```
## alternative hypothesis: true delta is less than 0
## sample estimates:
##      delta
## -0.4992864
##
##
## ###############################################
## # Augmented Dickey-Fuller Test Unit Root Test #
## ###############################################
##
## Test regression trend
##
##
## Call:
## lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -12819434   -544690    -32629    524047   8865013
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.623e+05  5.156e+04  10.905  < 2e-16 ***
## z.lag.1     -4.993e-01  2.724e-02 -18.328  < 2e-16 ***
## tt           5.811e+01  2.211e+01   2.629  0.00861 **
## z.diff.lag1 -1.215e-01  2.668e-02  -4.554 5.47e-06 ***
## z.diff.lag2 -1.241e-01  2.496e-02  -4.971 7.01e-07 ***
## z.diff.lag3 -4.686e-02  2.317e-02  -2.022  0.04324 *
## z.diff.lag4 -5.368e-02  2.047e-02  -2.622  0.00878 **
## z.diff.lag5 -9.256e-02  1.739e-02  -5.322 1.10e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1188000 on 3276 degrees of freedom
## Multiple R-squared:  0.3156, Adjusted R-squared:  0.3141
## F-statistic: 215.8 on 7 and 3276 DF,  p-value: < 2.2e-16
##
##
## Value of test-statistic is: -18.3281 111.975 167.9616
##
## Critical values for test statistics:
##       1pct  5pct 10pct
## tau3 -3.96 -3.41 -3.12
## phi2  6.09  4.68  4.03
## phi3  8.27  6.25  5.34
##
##
##  Phillips-Perron Unit Root Test
##
## data:  V_fracD
## Dickey-Fuller Z(alpha) = -2228.2, Truncation lag parameter = 9,
## p-value = 0.01
## alternative hypothesis: stationary
##
```

```
##
##  KPSS Test for Trend Stationarity
##
## data:  V_fracD
## KPSS Trend = 0.15688, Truncation lag parameter = 9, p-value =
## 0.04093
##
## [1] "-------------------------"
```

In this case, it barely didn't pass the KPSS test while it passed unit root tests. Then I tried d=0.81.

```
V_fracD=fracDiff(x=trainDat$V,d=0.81,tau=0.0001)
V_fracD=V_fracD[!is.na(V_fracD)]
tseries::adf.test(V_fracD)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  V_fracD
## Dickey-Fuller = -11.952, Lag order = 14, p-value = 0.01
## alternative hypothesis: stationary
```

```
CADFtest::CADFtest(V_fracD, type="trend", max.lag.y=5)
```

```
##
##  ADF test
##
## data:  V_fracD
## ADF(5) = -18.574, p-value < 2.2e-16
## alternative hypothesis: true delta is less than 0
## sample estimates:
##     delta
## -0.515962
```

```
summary(urca::ur.df(V_fracD, type="trend", lags=5))
```

```
##
## ###############################################
## # Augmented Dickey-Fuller Test Unit Root Test #
## ###############################################
##
## Test regression trend
##
##
## Call:
## lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -12828632   -540604    -31127    523889   8854012
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.329e+05  5.041e+04  10.571  < 2e-16 ***
## z.lag.1     -5.160e-01  2.778e-02 -18.574  < 2e-16 ***
## tt           5.413e+01  2.207e+01   2.453   0.0142 *
## z.diff.lag1 -1.151e-01  2.703e-02  -4.258 2.12e-05 ***
```

```
## z.diff.lag2 -1.192e-01  2.523e-02  -4.723 2.42e-06 ***
## z.diff.lag3 -4.369e-02  2.336e-02  -1.870   0.0615 .
## z.diff.lag4 -5.125e-02  2.057e-02  -2.491   0.0128 *
## z.diff.lag5 -9.117e-02  1.739e-02  -5.242 1.69e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1188000 on 3276 degrees of freedom
## Multiple R-squared:  0.3213, Adjusted R-squared:  0.3198
## F-statistic: 221.5 on 7 and 3276 DF,  p-value: < 2.2e-16
##
##
## Value of test-statistic is: -18.5738 114.9965 172.4938
##
## Critical values for test statistics:
##       1pct  5pct 10pct
## tau3 -3.96 -3.41 -3.12
## phi2  6.09  4.68  4.03
## phi3  8.27  6.25  5.34
```

```
tseries::pp.test(V_fracD)
```

```
##
##  Phillips-Perron Unit Root Test
##
## data:  V_fracD
## Dickey-Fuller Z(alpha) = -2257.7, Truncation lag parameter = 9,
## p-value = 0.01
## alternative hypothesis: stationary
```

```
tseries::kpss.test(V_fracD, null="Trend")
```

```
##
##  KPSS Test for Trend Stationarity
##
## data:  V_fracD
## KPSS Trend = 0.14117, Truncation lag parameter = 9, p-value =
## 0.05894
```

Finaly, it passed the stationarity test and d=0.81 is the appropriate value for fractionally differentiated volume.

Question3

```
C_fracD=fracDiff(x=dat$C,d=0.8,tau=0.0001)
V_fracD=fracDiff(x=dat$V,d=0.81,tau=0.0001)
```

```
library(devtools)
devtools::install_github("larryleihua/fmlr")
```

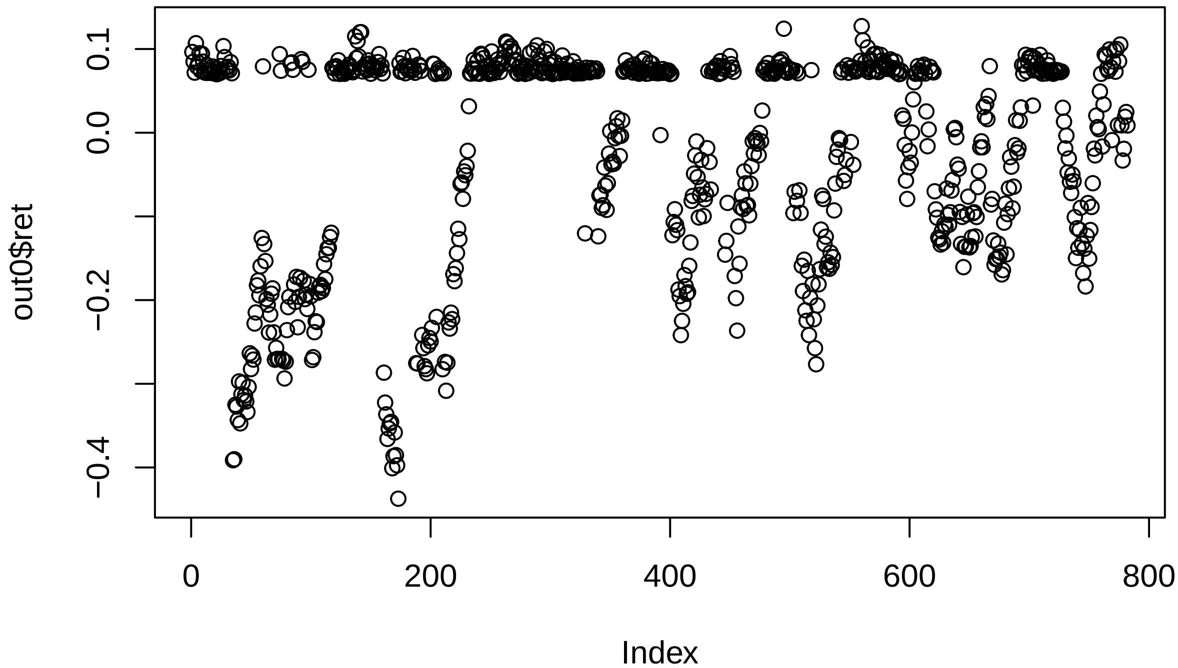
```
## Skipping install of 'fmlr' from a github remote, the SHA1 (c2b8ce53) has not changed since last insta
##   Use `force = TRUE` to force installation
```

```
#CUSUM filter on the raw closed prices series
i_CUSUM <- fmlr::istar_CUSUM(dat$C, h=200)
n_Event <- length(i_CUSUM)
```

```r
#triple barrier labeling method on the raw closed prices series
events <- data.frame(t0=i_CUSUM+1,t1 = i_CUSUM+200,trgt = rep(0.07, n_Event),side=rep(1,n_Event))
ptSl <- c(1,1)
out0 <- fmlr::label_meta(dat$C, events, ptSl, ex_vert = F)
table(out0$label)
```

```
##
##   0   1
## 374 408
```

```r
plot(out0$ret)
```



Question4

```r
#predictors
x <- dat$C
v <- dat$V
x_frac <- C_fracD
v_frac <- V_fracD

fMat0 <- t(sapply(1:nrow(out0),
                function(i){
                      i_range <- out0$t0Fea[i]:out0$t1Fea[i]
                      winTmp <- x[i_range]
                      winTmp_frac=x_frac[i_range]
                      C <- tail(winTmp,1)
                      V <- sum(v[i_range])
                      C_frac=tail(winTmp_frac,1)
                      V_frac=sum(v_frac[i_range])
                      return(c(C,V,C_frac,V_frac))
                    }))
```

```r
#change into dataframe
fMat0 <- data.frame(fMat0)
allset=cbind(fMat0,as.factor(out0$label))
names(allset) <- c("C", "V" ,"C_frac", "V_frac", "Y")

#Split the data into train and test data
i_CUSUM_train <- i_CUSUM[i_CUSUM <= nrow(trainDat)]
trainSet <- allset[1:length(i_CUSUM_train),]
testSet <- allset[(length(i_CUSUM_train)+1):nrow(allset),]
```

```r
library(randomForest)
```

```
## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:timeSeries':
##
##     outlier
```

```r
library(ROCR)
```

```
## Loading required package: gplots

##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess
```

```r
library(adabag)
```

```
## Loading required package: rpart

## Loading required package: caret

## Loading required package: lattice

## Loading required package: ggplot2

##
## Attaching package: 'ggplot2'

## The following object is masked from 'package:randomForest':
##
##     margin

## Loading required package: foreach

## Loading required package: doParallel

## Loading required package: iterators

## Loading required package: parallel
```

```r
library(xgboost)
library(rpart)
```

```r
#bagged classification trees when h=200 and trgt=0.07
set.seed(1)
bag <- randomForest(Y ~ C + V + C_frac + V_frac, data = trainSet, mtry = 4, importance = TRUE, ntrees =

#Predection and confusion table
prob_test <- predict(bag, newdata=testSet, type="prob")
table(testSet$Y, prob_test[,2] >= 0.5)
```

```
##
##     FALSE TRUE
##   0    24   40
##   1     7   39
```

There are more than 10 labels of 1 for the features bars obtained from the test set.

```r
#AUC for the test set
pred <- prediction(ifelse(prob_test[,2]>=0.5, 1, 0), testSet$Y)
auc <- performance(pred, measure = "auc")@y.values[[1]]
auc
```

```
## [1] 0.611413
```

AUC for the test set predicted by the bagged trees was 0.611413 and this achieved the goal.

```r
acc_perf <- performance(pred, measure = "acc")
acc_vec <- acc_perf@y.values[[1]]
acc <- acc_vec[max(which(acc_perf@x.values[[1]] >= 0.5))]
fmlr::acc_lucky(train_class = table(trainSet$Y),test_class = table(testSet$Y), my_acc = acc)
```

```
## $my_accuracy
## [1] 0.5727273
##
## $p_random_guess
## [1] 0.086
##
## $p_educated_guess
## [1] 0.056
##
## $mean_random_guess
## [1] 0.5010091
##
## $mean_educated_guess
## [1] 0.4938182
##
## $acc_majority_guess
## [1] 0.4181818
```

It turned out that candidate model with the tuned parameters could overperform various guesses.