Student ID__A0197899R__________ **Student Name __ZHENG HAO_________**

Midterm

Tuesday, 1 October 2019@7-10pm

1. Do bring your laptop for the midterm.
2. Download
   i) your individual data (named after your student ID) and
   ii) d-cdsALL.txt
   from the folder
   LumiNUS->files->Midterm
   Using wrong data will lead to 0 marks.
3. You are required to fill in the following table to summarize your results and place the table on the front page of your answer sheets.
4. Your answers and R code should be properly named following the format
   StudentID_studentName.docx or StudentID_studentName.pdf
   and StudentID_studentName.R.
5. Your answer sheets and R code should be submitted to the folder
   LumiNUS -> files -> Midterm Student Submission
6. **The submission folder will be closed punctually at 10:00pm** (LumiNUS clock). A 10-30% penalty will apply to late submission. Any late submission will get at least a 10% penalty. A late submission received on the next day, i.e. 2 October, will be given a 20% penalty. A submission later than 2 days, i.e. from 3 October onwards, will get a 30% penalty.
7. This is an open-book test. You are allowed to use any teaching materials and internet resources. *HOWEVER, this is an individual test! You shouldn't collude with any other individual, or plagiarize their work. Suspected collusion or plagiarism will be dealt with according to the NUS' assessment regulations.*
   --- Do not let other people see your answers.
   --- If other students ask for help, tell them to ask the lecturer.

	Student ID	**A0197899R**
Q1.	Type of stationarity	**Non-stationary**
	Fitted model	**~arma(1,1)+garch(1,1),con.dist='std'**
Q2.	Fitted model	**~arma(1,4)+garch(1,1)**
	Interval forecasts	**Step1 = [89.00989,89.02330] , step2 = [89.00659,89.02004]**
Q3	Fitted model	**~arma(5,5)+garch(1,1)**
	Fitted model with dummy	**~arma(0,5)+garch(1,1)**

Student ID__A0197899R__________ Student Name __ZHENG HAO_________

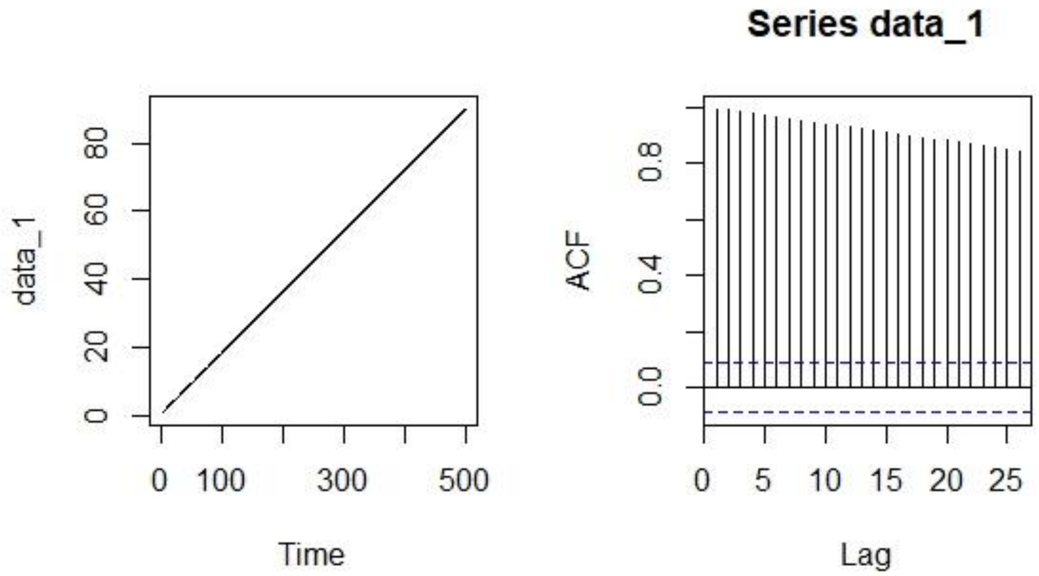
Analyze data and answer the following 3 questions with proper justifications.

Load your data using the following R code
read.table(filename, sep="\t", header=FALSE)

Question 1. Consider the first time series (the second column in your data), which is monthly sales of SDA – a FinTech company.

1. Plot the time series and ACF plots. Do you think the time series is stationary? Why or why not? What type of non-stationarity can you see in the series and how do you handle this type of non-stationarity? Do you observe any heteroskedasticity in the time series?

   Answer: the time series and acf plots are as the following. We can see the time series have obvious trend and the acf of it decrease very slowly. So I think it is not stationary. Then I will do the ADF and KPSS test.



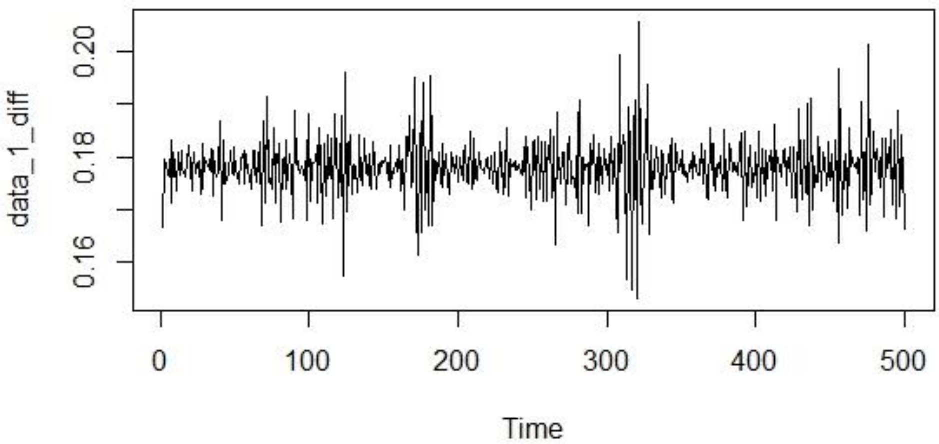
Student ID__A0197899R__________ Student Name __ZHENG HAO_________

```
        Augmented Dickey-Fuller Test

data:  data_1
Dickey-Fuller = -7.7034, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(data_1) : p-value smaller than printed p-value
> kpss.test(data_1)

        KPSS Test for Level Stationarity

data:  data_1
KPSS Level = 8.4317, Truncation lag parameter = 5, p-value = 0.01
```

Both ADF and KPSS reject the null hypothesis, to be conservative, and also with the information of the ts plot, let's assume the original data to be non-stationary.

And it is hard tell

So I choose to do one order difference of the data. And the diff(data) is:



It seems mean-reverting, we do the ADF and KPSS test again:

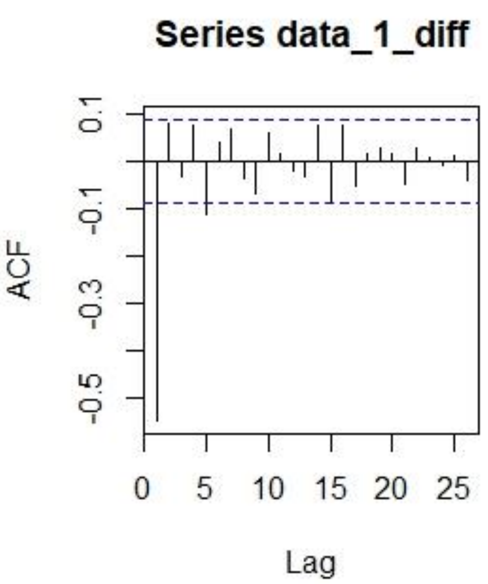
Student ID__A0197899R__________   Student Name __ZHENG HAO_________

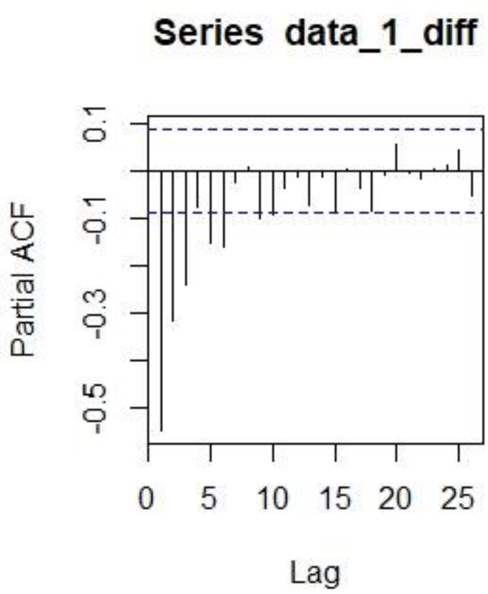
```
        Augmented Dickey-Fuller Test

data:  data_1_diff
Dickey-Fuller = -11.577, Lag order = 7, p-value =
0.01
alternative hypothesis: stationary

Warning message:
In adf.test(data_1_diff) : p-value smaller than printed p-
> kpss.test(data_1_diff)

        KPSS Test for Level Stationarity

data:  data_1_diff
KPSS Level = 0.017571, Truncation lag parameter = 5,
p-value = 0.1
```

We can see that the ADF and KPSS give the same conclusion that the diff_ts is stationary.
And Heteroskedastic is hard to tell from these two plots.

2. Propose an appropriate model for the sales data and justify your model assumptions with supporting evidence, including but not necessarily limited to autocorrelations, GARCH effect and distributional assumption.

   Answer: From question 1 we know that we accept the one order difference of original data is stationary, so we plot the acf and pacf of the diff(data):



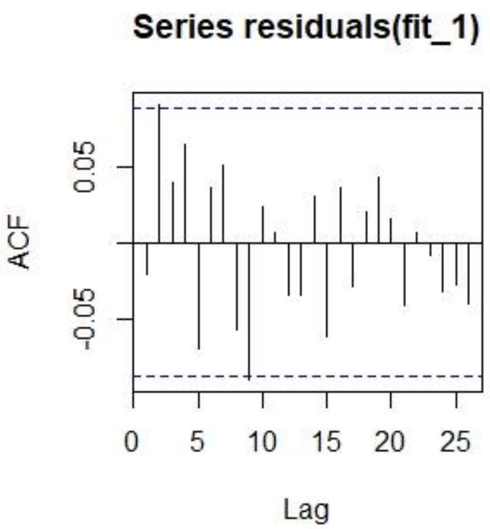


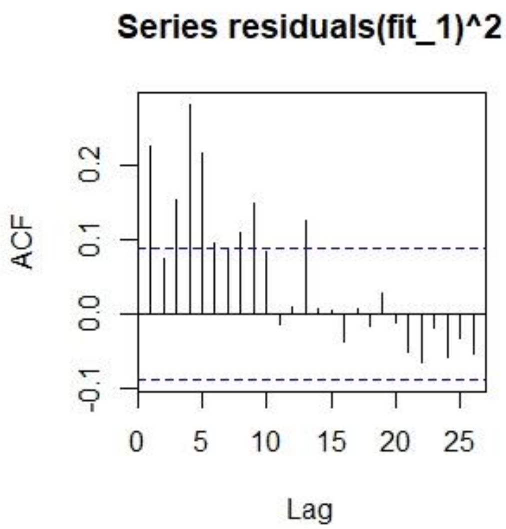
   We can see from the ACF that this data may fit MA(1) model, so we use

Student ID__A0197899R__________ Student Name __ZHENG HAO_________

arima(0,0,1) for the diff_data and we plot the ACF of residuals and residuals square.

```
fit_1 = arima(data_1_diff,order=c(0,0,1));
acf(residuals(fit_1))
acf(residuals(fit_1)^2)
```





I can see that ARIMA(0,1,1) may fit well for the data, however, the residuals have obvious Garch effect according the acf(residuals^2) plot. Thus I choose use Garch model to fit this data.

I choose three model as the following:

```
gfit_1 = garchFit(~arma(0,1)+garch(1,1),data=data_1_diff)
gfit_2 = garchFit(~arma(1,0)+garch(1,1),data=data_1_diff)
gfit_3 = garchFit(~arma(1,1)+garch(1,1),data=data_1_diff)
```

according to the AIC and BIC of each model, the gfit_3 model fit the best.

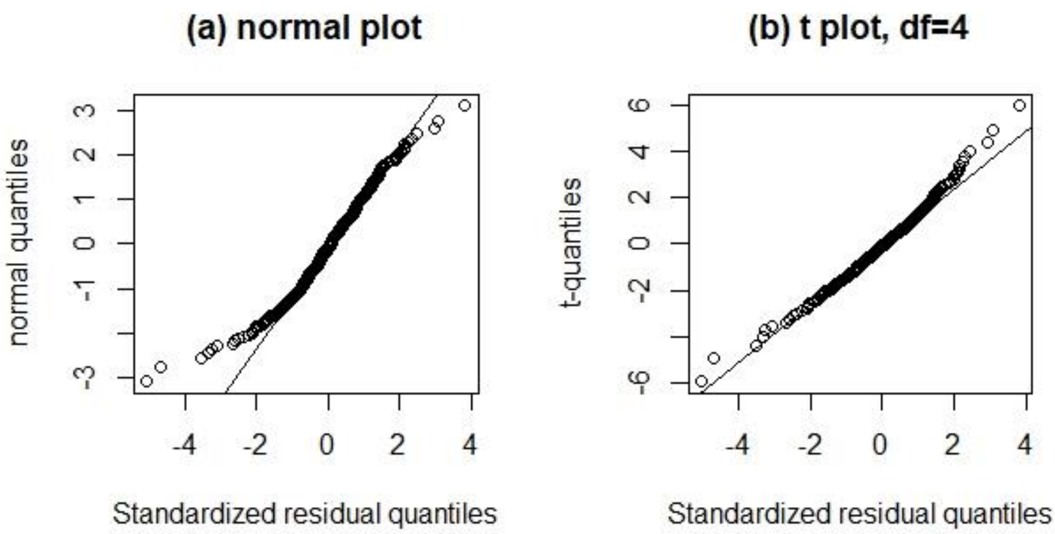
Student ID__A0197899R__________ Student Name __ZHENG HAO_________

```
Error Analysis:
        Estimate  Std. Error  t value Pr(>|t|)    
mu      1.816e-01          NA       NA       NA    
ar1    -2.040e-02   3.596e-05 -567.302  < 2e-16 ***
ma1    -1.000e+00          NA       NA       NA    
omega   2.047e-06   6.567e-07    3.116 0.001831 ** 
alpha1  1.951e-01   5.055e-02    3.859 0.000114 ***
beta1   6.938e-01   6.716e-02   10.330  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log Likelihood:
 2070.497    normalized:  4.149293 

Description:
 Tue Oct 01 19:29:14 2019 by user: zhenghaobaby
```

```
Standardised Residuals Tests:
                                Statistic p-Value     
 Jarque-Bera Test   R    Chi^2  216.3668  0           
 Shapiro-Wilk Test  R    W      0.9638317 9.768234e-10
 Ljung-Box Test     R    Q(10)  14.80296  0.1394123   
 Ljung-Box Test     R    Q(15)  20.92855  0.1391278   
 Ljung-Box Test     R    Q(20)  22.0783   0.3362772   
 Ljung-Box Test     R^2  Q(10)  8.146543  0.6145253   
 Ljung-Box Test     R^2  Q(15)  13.73447  0.5457551   
 Ljung-Box Test     R^2  Q(20)  22.64489  0.3065655   
 LM Arch Test       R    TR^2   9.191858  0.6864618   

Information Criterion Statistics:
      AIC       BIC       SIC      HQIC 
-8.274537 -8.223885 -8.274822 -8.254660 
```

Then we should see the residuals of this model.

Student ID__A0197899R__________ Student Name __ZHENG HAO_________



We can see that the residuals may more fit t-distribution rather than normal distribution. Thus we should change our assumption about the distribution.
Model:
gfit_4 = garchFit(~arma(1,1)+garch(1,1),data=data_1_diff,cond.dist = 'std')

3. Write down the fitted model for sales. Interpret the meaning of each coefficient and discuss significance.

Answer:

$$(r_t - 1.8166e^\wedge - 1) = 2.0562e^\wedge - 2\,(r_{t-1} - 1.8166e^\wedge - 1) + \varepsilon_t - \varepsilon_{t-1}$$

$$\varepsilon_t = \sqrt{h_t} * z_t \quad Zt \sim t(4.6281)$$

$$h_t = 1.4808e^{-6} + 1.8773\mathrm{e}^{-1}\varepsilon_{t-1}{}^2 + 0.74581 h_{t-1}$$

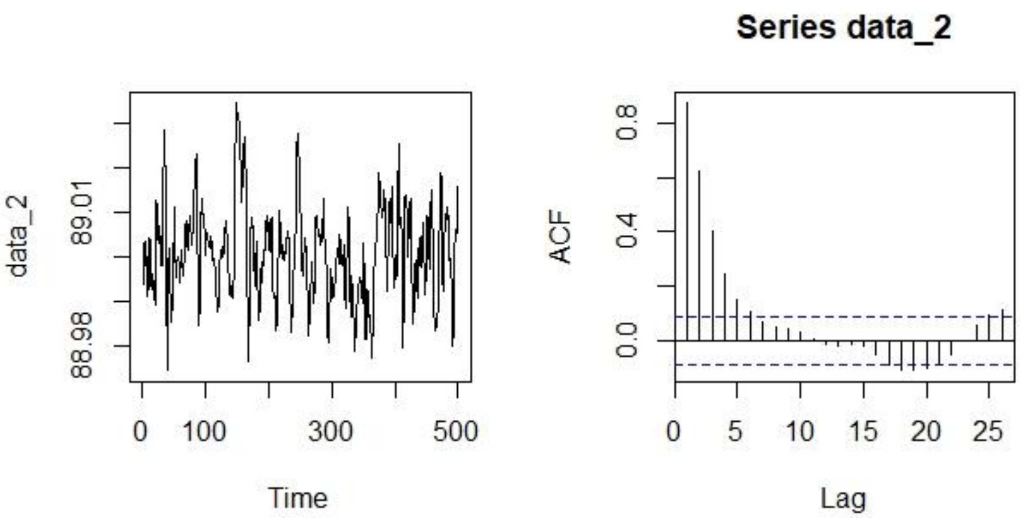
```
Call:
 garchFit(formula = ~arma(1, 1) + garch(1, 1), data = data_1_diff,
    cond.dist = "std")

Mean and Variance Equation:
 data ~ arma(1, 1) + garch(1, 1)
<environment: 0x00000227f70ec908>
 [data = data_1_diff]

Conditional Distribution:
 std

Coefficient(s):
         mu          ar1          ma1        omega       alpha1        beta1        shape
 1.8166e-01  -2.0562e-02  -1.0000e+00   1.4808e-06   1.8773e-01   7.4581e-01   4.6281e+00
```

```
Error Analysis:
        Estimate  Std. Error   t value Pr(>|t|)
mu      1.817e-01   3.900e-06 46576.823  < 2e-16 ***
ar1    -2.056e-02   3.602e-05  -570.926  < 2e-16 ***
ma1    -1.000e+00   2.116e-02   -47.261  < 2e-16 ***
omega   1.481e-06   6.283e-07     2.357  0.01843 *
alpha1  1.877e-01   5.712e-02     3.287  0.00101 **
beta1   7.458e-01   6.373e-02    11.703  < 2e-16 ***
shape   4.628e+00   9.736e-01     4.753    2e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The mu, ar1, ma1 are the coefficients for conditional mean and the omega, alpha1, beta1 are the coefficients for conditional variance. And the shape is the df of t-distribution. We can see from the picture that all the coefficients are significant according to 5% standard.

Question 2. Over years, SDA made gains in many forms of investing in real estate, bonds, etc. as displayed in the third column of your data.

4. Plot the time series and ACF plots. Do you think the time series is stationary? If yes, what type of non-stationarity can you see in the series and how do you handle this type of non-stationarity? Do you observe any heteroskedasticity in the time series?
   Answer: First I plot the time-series and ACF.



It seems that the time series are mean-reverting, I choose to do the ADF and KPSS test.:

Student ID__A0197899R__________ Student Name __ZHENG HAO_________

```
        Augmented Dickey-Fuller Test

data:  data_2
Dickey-Fuller = -5.9825, Lag order = 7,
p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(data_2) : p-value smaller than printed p-value
> kpss.test(data_2)

        KPSS Test for Level Stationarity

data:  data_2
KPSS Level = 0.19497, Truncation lag
parameter = 5, p-value = 0.1
```

Both ADF and KPSS test conclude that the time series are stationary. From the time series plots I think there exists **heteroskedasticity** because some time it has large variance.

5. Propose an appropriate model for the gains data. Write down the fitted model and report its AIC. Interpret the meaning of each coefficient and discuss significance.
   Answer: fit_2_g = garchFit(~arma(1,4)+garch(1,1),data=data_2)
   Model arma(1,4)+garch(1,1)

```
Error Analysis:
        Estimate  Std. Error   t value Pr(>|t|)
mu      2.425e+01   3.447e-04 70358.954  < 2e-16 ***
ar1     7.275e-01   3.405e-05 21366.270  < 2e-16 ***
ma1     1.000e+00   5.041e-02    19.836  < 2e-16 ***
ma2     2.078e-01   7.272e-02     2.857  0.00428 **
ma3     4.321e-02   7.363e-02     0.587  0.55731
ma4     3.541e-03   4.721e-02     0.075  0.94022
omega   1.124e-06   4.192e-07     2.682  0.00733 **
alpha1  5.319e-02   2.222e-02     2.394  0.01668 *
beta1   8.570e-01   4.258e-02    20.124  < 2e-16 ***
---
```

Standard AIC is

```
Information Criterion Statistics:
      AIC       BIC       SIC      HQIC
-8.446588 -8.370725 -8.447220 -8.416819
```

So the AIC IS

```
> -8.446588*length(data_2)
[1] -4223.294
>
```

The mu, ar1,ma1,ma2,ma3,ma4 are coefficients for conditional mean, and the omega, alpha1,beta1 are coefficients for conditional variance. The **ma3 and ma4** can is not significant according to the %5 standard and other coefficients are all significant.

6. Compute 1- and 2-step ahead forecasts and show their confidence interval.
   Answer:

```
  meanForecast   meanError standardDeviation
1     89.01659 0.003419152       0.003419152
2     89.01332 0.006830294       0.003429992
```

From the picture we can see the 1and 2 step ahead forecast and the interval is as the following:

```
> interval_low
[1] 89.00989 89.00659
> interval_high
[1] 89.02330 89.02004
>
```

Step1 = [89.00989,89.02330] , step2 = [89.00659,89.02004]

Question 3. Consider the daily CDS spreads (3-year maturity) of Allstate Insurance from January 01, 2004 to September 19, 2014. The period includes the financial crisis of 2008 so that the CDS spreads vary substantially. The data are in the file d-cdsALL.txt (column 2). Since the spreads are small, we consider the time series x_t = 100 × (spread3y). In addition, sample ACF of x_t shows strong persistence in the serial dependence. Therefore, we analyze the differenced series $y_t = (1 - B)x_t$.

7. Build a time series model for y_t. Write down the fitted model. [You may start with a model suggested by the auto.arima command in forecast package.]
   Answer: garchFit(~arma(5,5)+garch(1,1),data=data_3_diff)
   Model: arma(5,5)+garch(1,1)

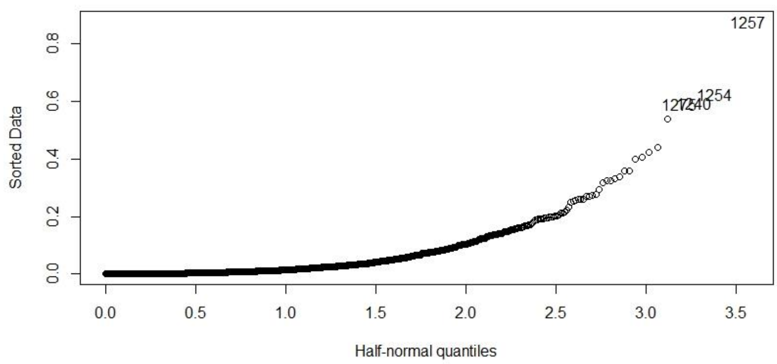
Student ID__A0197899R_________ **Student Name __ZHENG HAO________**

```
Error Analysis:
          Estimate  Std. Error  t value Pr(>|t|)
mu      -1.055e-06   3.788e-05   -0.028  0.97778
ar1      2.840e-01   1.147e-01    2.476  0.01330 *
ar2      3.764e-02   8.240e-02    0.457  0.64781
ar3     -9.285e-03   1.009e-01   -0.092  0.92669
ar4     -2.556e-01   1.036e-01   -2.467  0.01364 *
ar5      7.360e-01   7.498e-02    9.816  < 2e-16 ***
ma1     -2.913e-01   1.176e-01   -2.478  0.01321 *
ma2     -6.066e-04   7.704e-02   -0.008  0.99372
ma3     -2.999e-02   9.374e-02   -0.320  0.74906
ma4      3.294e-01   1.020e-01    3.230  0.00124 **
ma5     -7.428e-01   8.069e-02   -9.206  < 2e-16 ***
omega    3.270e-07   3.956e-10  826.557  < 2e-16 ***
alpha1   1.725e-01   1.457e-02   11.838  < 2e-16 ***
beta1    8.629e-01   8.492e-03  101.616  < 2e-16 ***
---
```

AIC:

```
Information Criterion Statistics:
      AIC       BIC       SIC      HQIC
-5.314800 -5.284810 -5.314851 -5.303968

> -5.3148*length(data_3_diff)
[1] -14700.74
>
```

8. To improve the fit, identify sequentially the largest four outliers of the fitted model. Write down the fitted model with the four largest outliers included as dummy.
   Answer: First, use the halfnorm function to find the outliners



From this picture I can know the index of the largest four outliners is 1257,1254,1240,1275, Thus make this four to be dummy. Then we fit the data again and get the model as the following:

Student ID__A0197899R__________ Student Name __ZHENG HAO_________

fit_3_g_mend = garchFit(~arma(0,5)+garch(1,1),data=data_3_diff_mend
Model: arma(0,5)+garch(1,1)

```
Information Criterion Statistics:
      AIC       BIC       SIC      HQIC
-5.330328 -5.311048 -5.330349 -5.323364

> -5.330328*length(data_3_diff_mend)
[1] -14743.69
>
```

And we can see the AIC significantly decrease by setting the largest four as dummy.

9. Let a_t be the residuals of the model in question (8) and ρ_i be the lag-i ACF of a_t. Test $H_0 : \rho_1 = \cdots = \rho_{10} = 0$ versus Ha : $\rho_i \neq 0$ for some $1 \leq i \leq 10$. Draw your conclusion.

Answer:
at = residuals(fit_3_g)
acf(at)
Box.test(at,lag=10,type="Ljung-Box")

at_std = residuals(fit_3_g)/fit_3_g@sigma.t
acf(at_std)
Box.test(at_std,lag=10,type="Ljung-Box")

```
        Box-Ljung test

data:  at
X-squared = 64.827, df = 10, p-value = 4.375e-10
```

We can see for residuals, the p-value is very small and we can reject the H_0.

```
        Box-Ljung test

data:  at_std
X-squared = 8.223, df = 10, p-value = 0.6071
```

However, for the standard residuals, we accept the H_0