Sofia Tsaklidou

Time Series and Forecasting Methods

Assignment: Time Series Analysis of JP Morgan US Fund

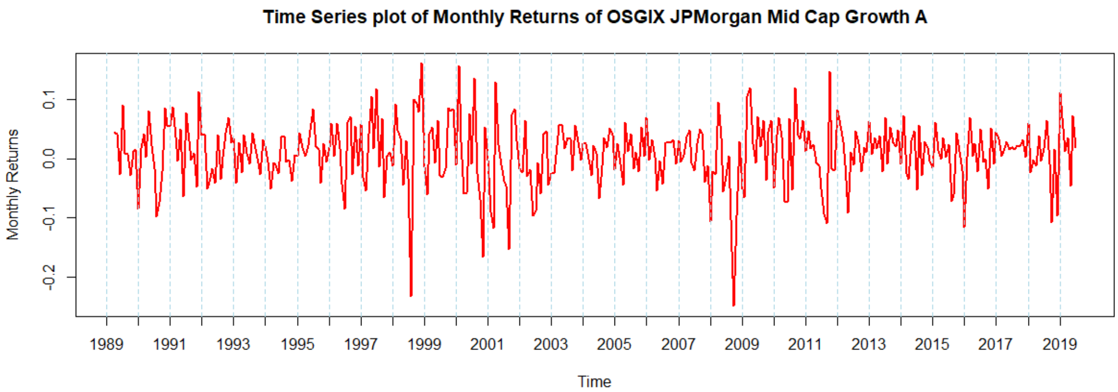
Supervisor: Dr. Vrontos

2021

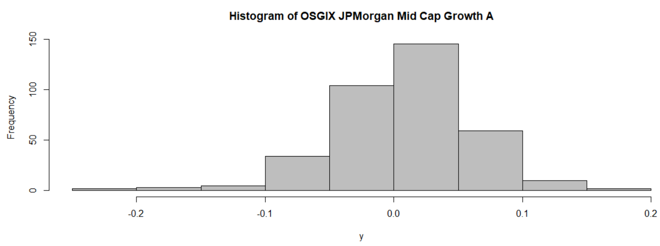
In this assignment we will perform a time series analysis on JP Morgan US Fund.
We will use the *OSGIX JPMorgan Mid Cap Growth A* as a dependent variable and the below as the independent variables:
$x_1 = Mkt - Rf, x_2 = SMB, x_3 = HML, x4 = RMW, x_5 = CMA, x_6 = MOM$
The starting date of dataset is August 1987 until July 2019. However, since we observe that there are null values from August 198 until April 1989 for this specific dependent variable we will start the analysis from April 1989 until July 2019.

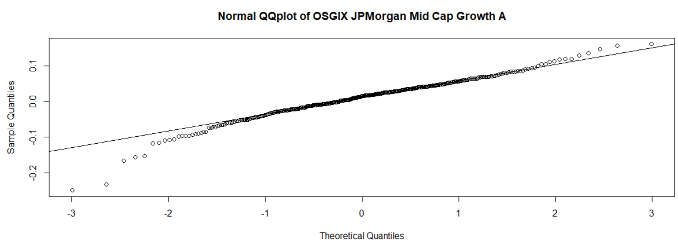
Data exploration

To start with, we will try to build time series models on some dependent variable OSGIX (AR, MA, ARMA). We will first test the hypothesis that the series of OSGIX is stationary over time using unit root testing along with other tests.



From the above plot we can observe that the mean of the data is around zero, specifically between -0.1 and 0.1, and there is no trend. However, there are a few peaks during random time periods. We can assume from the above plot that the series seems to be stationary.
Now, we will proceed with the histogram and the QQ$-$plot.





From the above histogram we can observe that there is a slight left skewness in the distribution of the data and the data appear to roughly follow a zero-mean normal distribution. In addition we observe that there is a slight left fat tail in the distribution.
In this case, we will proceed with the Shapiro-Wilk normality test.

```
Shapiro-Wilk normality test

data:  y
W = 0.96645, p-value = 2.019e-07
```

Since p-value = 2.019e-07, for $a = 0.05$ significance level, we reject the null hypothesis so we assume that the data is not normally distributed.

Now we can proceed with the augmented Dickey-Fuller test of unit root based on Random Walk with Drift, since we can do so as there was not any trend observed in the previous plots.

```
###############################################################
# Augmented Dickey-Fuller Test Unit Root / Cointegration Test #
###############################################################

The value of the test statistic is: -17.2711 149.146

###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression drift


Call:
lm(formula = z.diff ~ z.lag.1 + 1)

Residuals:
      Min        1Q    Median        3Q       Max
-0.243243 -0.030247  0.003265  0.033924  0.150493

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.008380   0.002875   2.914  0.00379 **
z.lag.1     -0.904326   0.052361 -17.271  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05399 on 361 degrees of freedom
Multiple R-squared:  0.4524,Adjusted R-squared:  0.4509
F-statistic: 298.3 on 1 and 361 DF,  p-value: < 2.2e-16


Value of test-statistic is: -17.2711 149.146

Critical values for test statistics:
      1pct  5pct 10pct
tau2 -3.44 -2.87 -2.57
phi1  6.47  4.61  3.79
```

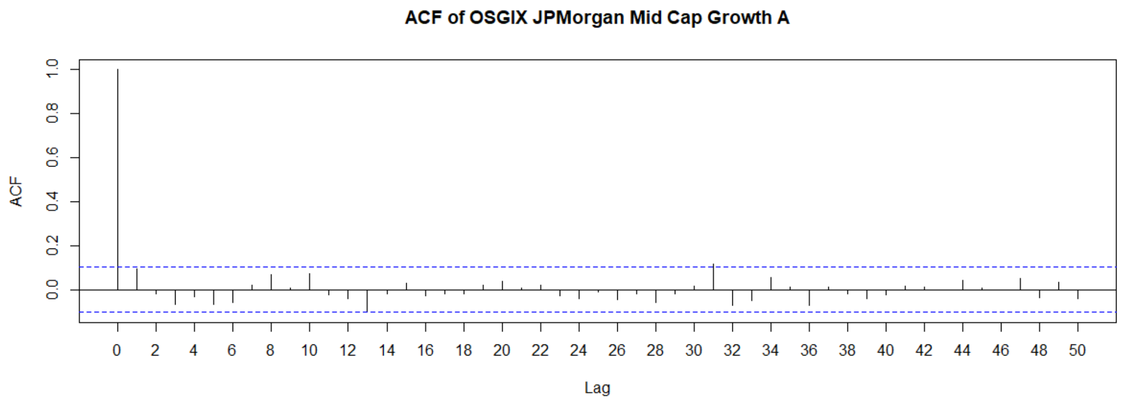
Let assume that the initial model is $Y_t = \mu + \rho Y_{t-1} + \epsilon_t$ and by applying reparametrization we get $\Delta Y = \mu + \beta Y_{t-1} + \epsilon_t$. So,

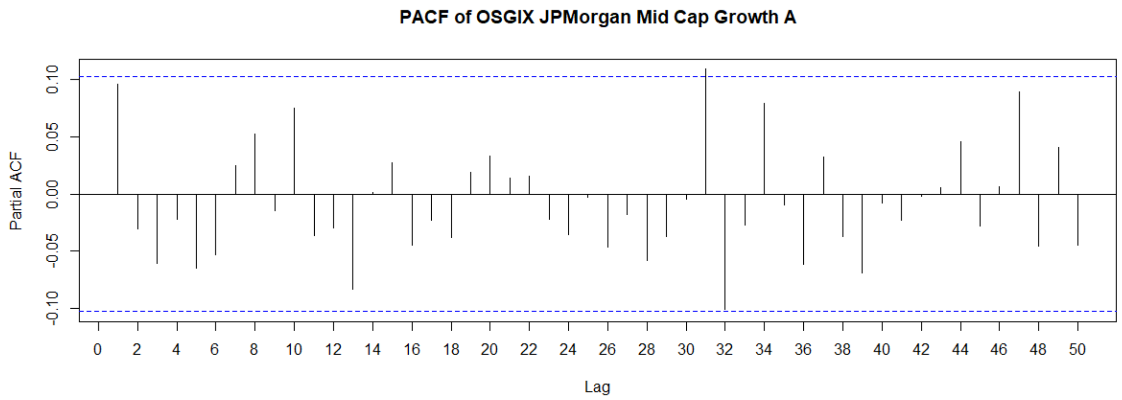
$$\text{H}_0\text{: } \beta = 0$$
$$\text{H}_a\text{: } \beta \neq 0$$

From the above results we observe that $\hat{\mu} = 0.008380$ and the $\hat{\beta} = -0.904326$. The value of test-statistic is: -17.2711 whereas the critical values are $\{-3.44, -2.87, -2.57\}$ for a $\{0.01, 0.05, 0.1\}$ significance levels respectively. This means that we reject the null hypothesis that the model has a unit root.
Now, we can proceed with the the Box-Jenkins methodology.

1.Identification step

To begin with, we will plot the ACF and the PACF of OSGIX.





It is known that thhe ACF and PACF plots should be considered together to define the process. For the AR process, we expect that the ACF plot will gradually decrease and simultaneously the PACF should have a sharp drop after p significant lags. To define a MA process, we expect the opposite from the ACF and PACF plots, meaning that: the ACF should show a sharp drop after a certain q number of lags while PACF should show a geometric or gradual decreasing trend.

On the above plots we observe that the fifty lag observations are located within the boundaries of significance, except for the lag 31 of PACF, where we observe that there is a slight violation. We have some indications of non-statistical significant autocorrelations regarding the values of the time series, however we will proceed with the Box-Pierce test for autocorrelations.

$$H_0\colon\ \rho_1 = \rho_2 = ... = \rho_{50} = 0$$
$$H_a\colon\ \rho_i \neq 0 \text{ for at least one } i \leq 50$$

```
Box-Pierce test

data:  y
X-squared = 36.767, df = 50, p-value = 0.9183
```

From the above, since p-value = 0.9183, we do not reject the null hypothesis for $a = 0.05$ significance level which means that $\rho_1 = \rho_2 = ... = \rho_{50} = 0$. In fact, now we can clearly state that we do not have statistically significant autocorrelations for $a = 0.05$ in any lag for OSGIX.

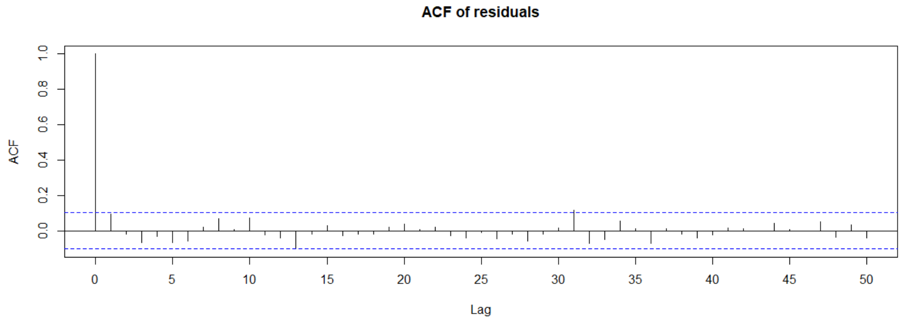
2.Estimation step

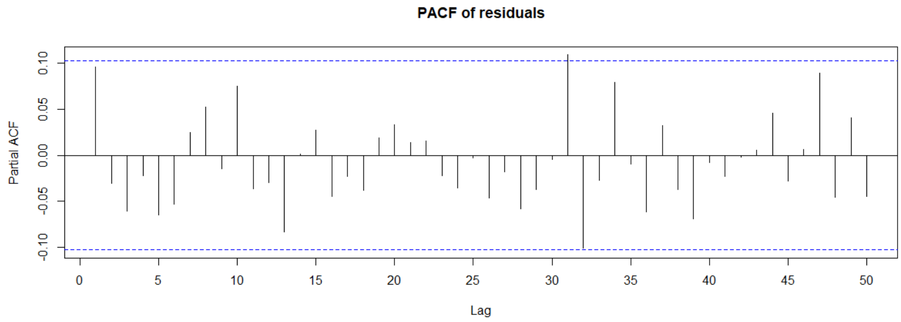
Since we do not have statistically significant autocorrelations for $a = 0.05$ in any lag for OSGIX, we will proceed withe the below model:

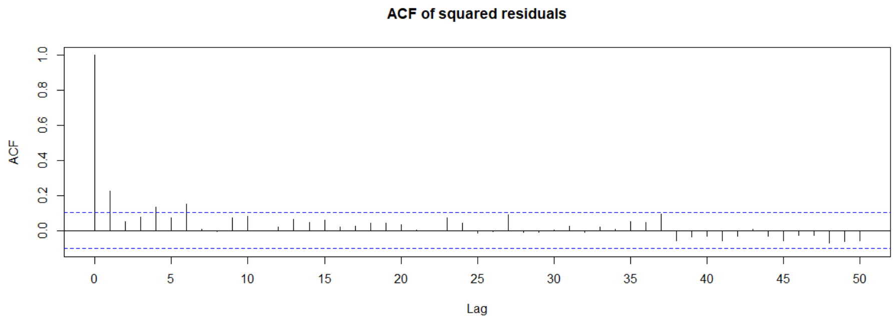
$$Y_t = \mu + \epsilon_t$$

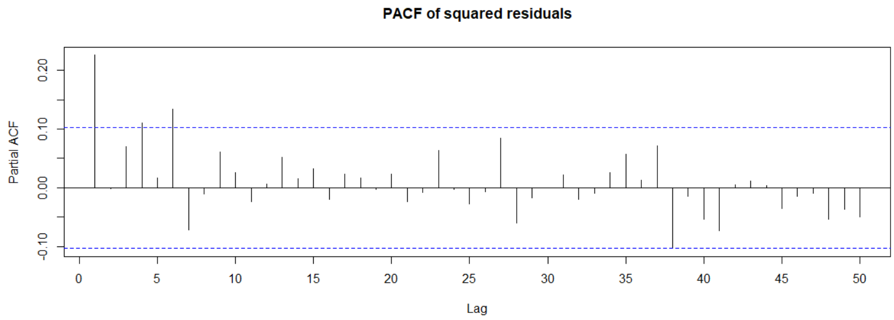
```
Call:
arima(x = OSGIX, order = c(0, 0, 0))

Coefficients:
      intercept
         0.0056
s.e.     0.0023

sigma^2 estimated as 0.001412:  log likelihood = 491.7,  aic = -979.41
```

3. Diagnostic plots









From the above plots can state that we should probably use a GARCH model. It is known that the ACF plot of squared residuals gives us information about the appropriate GARCH terms, whereas the PACF plot gives us information about the appropriate ARCH terms.

We can observe a slightly statistically significant autocorrelations for $a = 0.05$ in lags 1, 4, 6 of squared residuals and a slightly statistically significant partial autocorrelations for $a = 0.05$ in lags 1, 4 and 6.

This means that there is probably a violation of heteroscedasticity of the observed data.

Heteroscedasticity issue

```
Box-Ljung test

data:  residuals^2
X-squared = 74.181, df = 50, p-value = 0.0148
```

We observe that the p-value = 0.0148, so this means that we reject the null hypothesis. In other words, there is a $\rho_i \neq 0$.

Developing appropriate model

Having said the above, we will built an ARCH(1) model and a GARCH(1,1) and we will check which one fits better based on the AIC criterion.

ARCH(1)

```
Title:
 GARCH Modelling

Call:
 garchFit(formula = ~garch(1, 0), data = y, trace = F)

Mean and Variance Equation:
 data ~ garch(1, 0)
<environment: 0x000001fadc1d9e40>
 [data = y]

Conditional Distribution:
 norm

Coefficient(s):
       mu      omega     alpha1
0.0129388  0.0021041  0.2890552

Std. Errors:
 based on Hessian

Error Analysis:
        Estimate  Std. Error  t value Pr(>|t|)
mu     0.0129388   0.0027155    4.765 1.89e-06 ***
omega  0.0021041   0.0002229    9.438  < 2e-16 ***
alpha1 0.2890552   0.0953026    3.033  0.00242 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log Likelihood:
 558.5041    normalized:  1.534352

Description:
 Sun Nov 07 14:05:32 2021 by user: sofia


Standardised Residuals Tests:
                                Statistic p-Value
 Jarque-Bera Test   R    Chi^2  57.66983  2.999823e-13
 Shapiro-Wilk Test  R    W      0.978392  2.853769e-05
 Ljung-Box Test     R    Q(10)  10.74748  0.3775276
 Ljung-Box Test     R    Q(15)  16.82526  0.3294171
 Ljung-Box Test     R    Q(20)  18.14795  0.5776627
 Ljung-Box Test     R^2  Q(10)  20.91129  0.0217207
 Ljung-Box Test     R^2  Q(15)  26.41649  0.03386634
 Ljung-Box Test     R^2  Q(20)  32.09888  0.04225605
 LM Arch Test       R    TR^2   15.80833  0.2001733

Information Criterion Statistics:
      AIC       BIC       SIC      HQIC
-3.052220 -3.020101 -3.052354 -3.039454
```

So from the above we can conclude the below:

Proposed model: $Y_t = 0.0129388 + \epsilon_t$, $\sigma^2 = 0.0021041 + 0.2890552\epsilon_{t-1}^2$

Standardised Residuals Tests:

- **Jarque-Bera Test for Residuals**: We observe that the p-Value = 2.999823e-13. This means that we reject the null hypothesis at $a = 0.05$ significance level. In other words, t there is a violation in the normality of the Residuals.
- **Shapiro-Wilk Test for Residuals**: We observe that the p-Value = 2.853769e-05. This means that we reject the null hypothesis at $a = 0.05$ significance level. In other words, there is a violation in the normality of the Residuals.
- **Ljung-Box Test for Residuals**: There are three Ljung-Box tests for the Residuals and we observe that all the three p-Values are greater than $a = 0.05$ significance level. This means that we do not reject the null hypothesis, so there is no statistically significant autocorrelations in lags 10, 15 and 20.
- **Ljung-Box Test for squared Residuals**: There are three Ljung-Box tests for the squared Residuals and we observe that all the three p-Values are less than $a = 0.05$ significance level. This means that we reject the null hypothesis in lags 10, 15 and 20.
- **LM Arch Test for Residuals**: The ARCH-LM test is the standard test to detect autoregressive conditional heteroscedasticity. We observe that the p-Value = 0.2001733. This means that we do not reject the null hypothesis at $a = 0.05$ significance level, therefore the residuals are heteroscedastic.

GARCH(1, 1)

```
Title:
 GARCH Modelling

Call:
 garchFit(formula = ~garch(1, 1), data = y, trace = F)

Mean and Variance Equation:
 data ~ garch(1, 1)
<environment: 0x000001fadbe406d8>
 [data = y]

Conditional Distribution:
 norm

Coefficient(s):
        mu       omega      alpha1       beta1
0.01056794  0.00017365  0.15819631  0.79174581

Std. Errors:
 based on Hessian
```

```
Error Analysis:
        Estimate  Std. Error  t value Pr(>|t|)
mu     1.057e-02   2.484e-03    4.255 2.09e-05 ***
omega  1.736e-04   8.811e-05    1.971  0.04874 *
alpha1 1.582e-01   5.114e-02    3.094  0.00198 **
beta1  7.917e-01   6.164e-02   12.844  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log Likelihood:
 569.3302    normalized:  1.564094

Description:
 Sun Nov 07 14:44:52 2021 by user: sofia


Standardised Residuals Tests:
                                Statistic p-Value
 Jarque-Bera Test   R    Chi^2  89.06871  0
 Shapiro-Wilk Test  R    W      0.9688967 5.106265e-07
 Ljung-Box Test     R    Q(10)  12.8709   0.2309797
 Ljung-Box Test     R    Q(15)  17.54966  0.2870802
 Ljung-Box Test     R    Q(20)  20.20024  0.4454684
 Ljung-Box Test     R^2  Q(10)  2.593138  0.9894479
 Ljung-Box Test     R^2  Q(15)  4.59588   0.9950302
 Ljung-Box Test     R^2  Q(20)  6.279072  0.9984643
 LM Arch Test       R    TR^2   3.066375  0.995057

Information Criterion Statistics:
      AIC       BIC       SIC      HQIC
-3.106210 -3.063384 -3.106448 -3.089189
```

So from the above we can conclude the below:

Proposed model: $Y_t = 1.057e^{-02} + \epsilon_t$, $\sigma^2 = 1.736e^{-04} + 1.582e^{-01}\epsilon_{t-1}^2 + 7.917e^{-0}\sigma_{t-1}^2$

Standardised Residuals Tests:

- **Jarque-Bera Test for Residuals**: We observe that the p-Value = 0. This means that we reject the null hypothesis at $a = 0.05$ significance level. In other words, there is a violation in the normality of the Residuals.

- **Shapiro-Wilk Test for Residuals**: We observe that the p-Value = $5.106265e^{-07}$. This means that we reject the null hypothesis at $a = 0.05$ significance level. In other words, there is a violation in the normality of the Residuals.

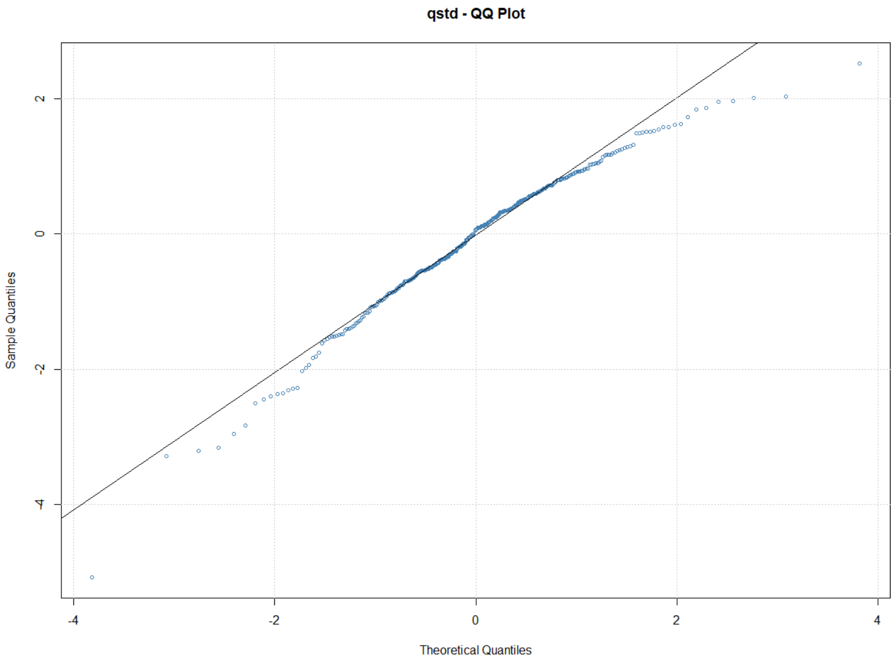
- **Ljung-Box Test for Residuals**: There are three Ljung-Box tests for the Residuals and we observe that all the three p-Values are greater than $a = 0.05$ significance level. This means that we do not reject the null hypothesis, so there is no statistically significant autocorrelations in lags 10, 15 and 20.

- **Ljung-Box Test for squared Residuals**: There are three Ljung-Box tests for the squared Residuals and we observe that all the three p-Values are greater than $a = 0.05$ significance level. This means that we do not reject the null hypothesis in lags 10, 15 and 20.

- **LM Arch Test for Residuals**: The ARCH-LM test is the standard test to detect autoregressive conditional heteroscedasticity. We observe that the p-Value = 0.995057. This means that we do not reject the null hypothesis at $a = 0.05$ significance level, therefore the residuals are heteroscedastic.

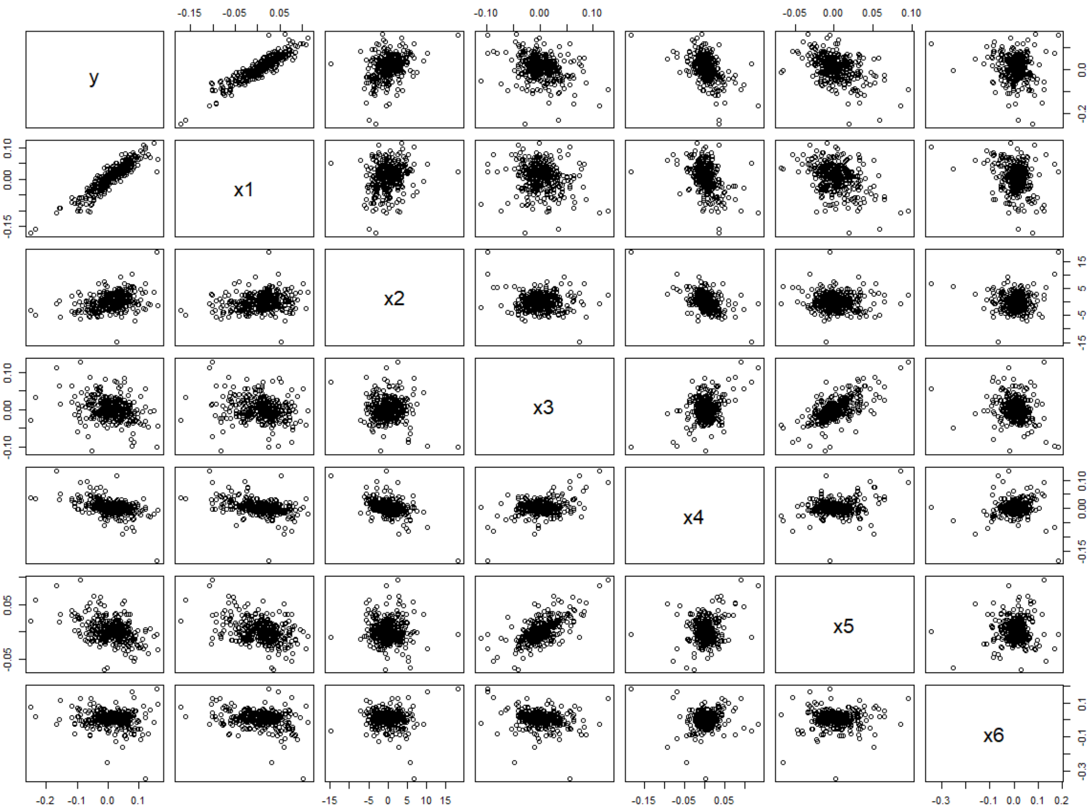
Taking into consideration the two models ARCH(1) and GARCH(1,1) we observe that the AIC is -3.052220 and -3.106210 respectively. Since the GARCH(1,1) has a lower AIC will we proceed with model.

In addition, running the Jarque-Bera and Shapiro-Wilk Test for Residual we have observed a slight violation the best scenario is to use a Student-t distribution.



Regression analysis

To begin with, we will start with a scatterplot of variables' correlation.



	y	x1	x2	x3	x4	x5	x6
y	1.00	0.92	0.37	-0.29	-0.47	-0.43	-0.12
x1	0.92	1.00	0.21	-0.17	-0.40	-0.38	-0.25
x2	0.37	0.21	1.00	-0.10	-0.46	-0.06	0.00
x3	-0.29	-0.17	-0.10	1.00	0.35	0.65	-0.20
x4	-0.47	-0.40	-0.46	0.35	1.00	0.22	0.08
x5	-0.43	-0.38	-0.06	0.65	0.22	1.00	0.03
x6	-0.12	-0.25	0.00	-0.20	0.08	0.03	1.00

n= 364

P

	y	x1	x2	x3	x4	x5	x6
y		0.0000	0.0000	0.0000	0.0000	0.0000	0.0235
x1	0.0000		0.0000	0.0010	0.0000	0.0000	0.0000
x2	0.0000	0.0000		0.0619	0.0000	0.2842	0.9286
x3	0.0000	0.0010	0.0619		0.0000	0.0000	0.0001
x4	0.0000	0.0000	0.0000	0.0000		0.0000	0.1443
x5	0.0000	0.0000	0.2842	0.0000	0.0000		0.5066
x6	0.0235	0.0000	0.9286	0.0001	0.1443	0.5066	

From the previous page we observe that there is a statistically significant correlation among the variables Y and x1, x2, x3, x4, x5, x6 at $a = 0.05$ significance level. This means that the variables are independent.

Let's proceed with the regression model:

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6)

Residuals:
      Min        1Q    Median        3Q       Max
-0.069669 -0.010662  0.000553  0.009312  0.077207

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0014722  0.0009946   1.480   0.1397
x1           1.1074995  0.0266650  41.534  < 2e-16 ***
x2           0.0032032  0.0003498   9.156  < 2e-16 ***
x3          -0.1167858  0.0470280  -2.483   0.0135 *
x4          -0.0123838  0.0469052  -0.264   0.7919
x5          -0.1491719  0.0663786  -2.247   0.0252 *
x6           0.1013989  0.0215202   4.712 3.52e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01774 on 357 degrees of freedom
Multiple R-squared:  0.8943,Adjusted R-squared:  0.8926
F-statistic: 503.6 on 6 and 357 DF,  p-value: < 2.2e-16
```

From the above we observe that the Adjusted R-squared = 0.8926, which is really good. We observe also that all coefficients of the x variables, except for x4, are statistically significant. We will use the stepwise elimination methods in order to insert or remove independent variables from our model according to BIC.

Stepwise elimination method

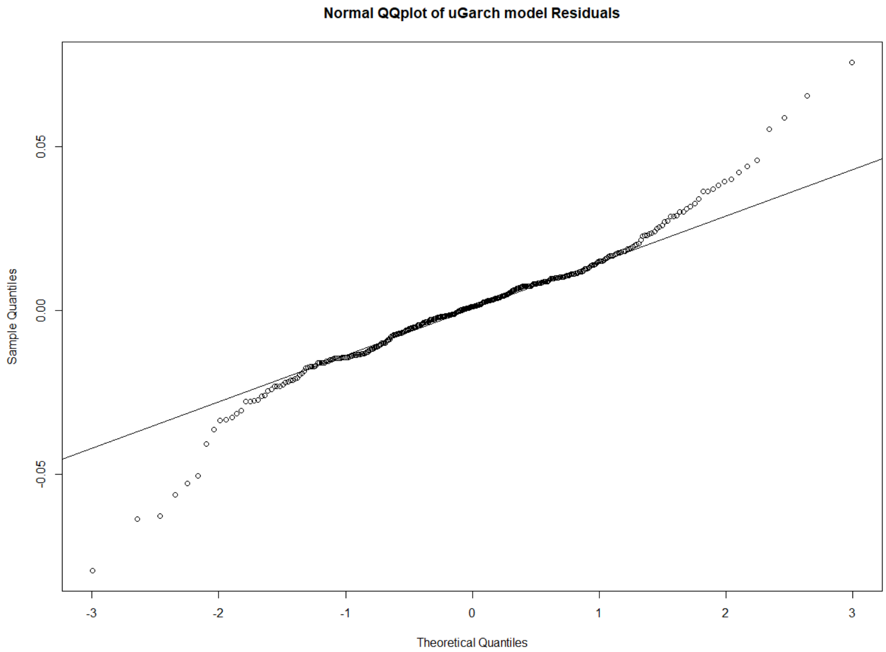
```
  Step Df    Deviance Resid. Df Resid. Dev       AIC
1      NA          NA       363  1.0631543 -2122.273
2 + x1 -1 0.892201831       362  0.1709525 -2785.523
3 + x2 -1 0.035785200       361  0.1351673 -2869.016
4 + x3 -1 0.014993442       360  0.1201739 -2909.813
5 + x6 -1 0.006253588       359  0.1139203 -2927.265
6 + x5 -1 0.001569866       358  0.1123504 -2930.316


Call:
lm(formula = y ~ x1 + x2 + x3 + x6 + x5)

Coefficients:
(Intercept)           x1           x2           x3           x6           x5
   0.001413     1.109789     0.003242    -0.121056     0.100834    -0.146394
```

The process starts with an empty model. In the first step it adds 1 because it offers the greater reduction of BIC in comparison to the other variables. Then it adds x2 e.t.c.

Combining everything we have done so far, we will proceed with a multiple regression model, with external regressors and the proposed Garch(1,1).

```
*---------------------------------*
*          GARCH Model Fit        *
*---------------------------------*

Conditional Variance Dynamics
-----------------------------------
GARCH Model : sGARCH(1,1)
Mean Model : ARFIMA(0,0,0)
Distribution : norm

Optimal Parameters
------------------------------------
        Estimate  Std. Error  t value Pr(>|t|)
mu      0.001332    0.000796   1.6722 0.094476
mxreg1  1.042074    0.021211  49.1284 0.000000
mxreg2  0.003633    0.000307  11.8506 0.000000
mxreg3 -0.094460    0.042818  -2.2061 0.027380
mxreg4 -0.050626    0.044699  -1.1326 0.257384
mxreg5 -0.322848    0.057979  -5.5684 0.000000
mxreg6  0.115617    0.020500   5.6398 0.000000
omega   0.000029    0.000013   2.1983 0.027928
alpha1  0.242350    0.073371   3.3031 0.000956
beta1   0.661919    0.098581   6.7144 0.000000

Robust Standard Errors:
        Estimate  Std. Error  t value Pr(>|t|)
mu      0.001332    0.000893  1.49206 0.135684
mxreg1  1.042074    0.022669 45.96893 0.000000
mxreg2  0.003633    0.000354 10.25949 0.000000
mxreg3 -0.094460    0.056369 -1.67575 0.093787
mxreg4 -0.050626    0.050731 -0.99792 0.318320
mxreg5 -0.322848    0.067183 -4.80549 0.000002
mxreg6  0.115617    0.026837  4.30819 0.000016
omega   0.000029    0.000015  2.02134 0.043245
alpha1  0.242350    0.083708  2.89519 0.003789
beta1   0.661919    0.113562  5.82868 0.000000

LogLikelihood : 999.8167

Information Criteria
------------------------------------

Akaike       -5.4386
Bayes        -5.3315
Shibata      -5.4400
Hannan-Quinn -5.3960
```

From the above we also observe that x3 , x4 are possibly statistically insignificant, since their corresponding p-values are greater than the $a = 0.5$ significance level.



From the above plot we observe that there are fat tails so the normality have been violated.
In addition, running the Jarque-Bera and Shapiro-Wilk Test for Residual we have observed a violation and as said also in the begging the best scenario is to use a Student-t distribution.

```
Jarque Bera Test

data:  residuals(model_res)
X-squared = 140.49, df = 2, p-value < 2.2e-16
```

AIC and BIC information criteria

Excluding x3

```
Information Criteria
------------------------------------

Akaike       -5.4285
Bayes        -5.3214
Shibata      -5.4299
Hannan-Quinn -5.3859
```

Excluding x4

```
Information Criteria
------------------------------------

Akaike       -5.4419
Bayes        -5.3348
Shibata      -5.4433
Hannan-Quinn -5.3993
```

Excluding x3 and x4

```
Information Criteria
------------------------------------

Akaike       -5.4276
Bayes        -5.3312
Shibata      -5.4288
Hannan-Quinn -5.3893
```

From the above we observe that the best AIC score is obtained when we removed only the x4 regressor. So, the proposed model is:

$$Y_t = \gamma_0 + \gamma_1 Y_{1,t} + \gamma_2 Y_{2,t} + \gamma_3 Y_{3,t} + \gamma_5 Y_{5,t} + \gamma_6 Y_{6,t}$$

$$\epsilon_t | \Phi t - 1 \sim N(0, \sigma_t^2)$$

$$\sigma_t^2 = a_0 + a\epsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2$$