Data analysis- (Tesla stock)

Introduction and Background

We will try to forecast daily closing price of Tesla shares using data spanning from 2018-09-12 to 2022-05-13. We will be using ARIMA GARCH modelling to make the forecast, by fitting several models and the select the best model out of the rest.

We will also try to fit a regression model to the data in order to see if our variable are influencing Tesla closing price.

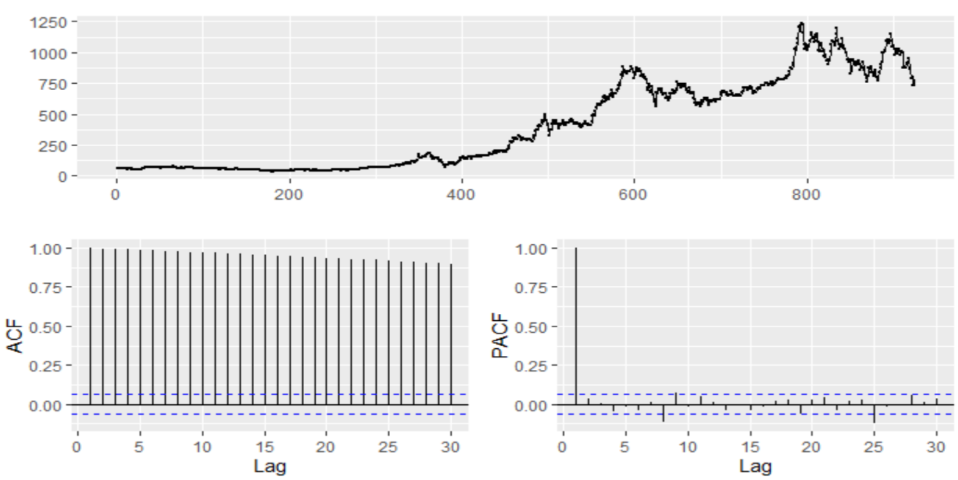
Method of choosing Model:

We first need to transform the data into a time series object. With this we can create a time series model, Autocorrelation Function, and Partial Autocorrelation Function to determine the autoregressive value (p), difference, and moving average. We will first look at the time series model and see if it is stationary or not, if it is stationary, we do need to use a difference function. With this information we can create a few different test models and calculate their Akaike Information Criterion correlated (AICc), and those with the lowest AICc will fit the data the best. We will also use the Auto ARIMA function and to see how its AICc compares to the ones we tested. Finally, we will use the Ljung- Box-lack of fit test to calculate the p-value for the model, if it is greater than 0.05 then we can conclude it fits the data well

We start by plotting the time series and observe that the data have a non -constant mean and variance, showing non-stationarity of the data. We need to log the data to in order to make it variance constant and take the first seasonal difference of it to solve the issue of non-stationarity. Afterwards, we plotted the ACF and PACF of the logged and differenced data in order to have an idea of how our model would look like. Our aim now is to find an appropriate ARIMA model based on the ACF and PACF

Fitting ARIMA model:

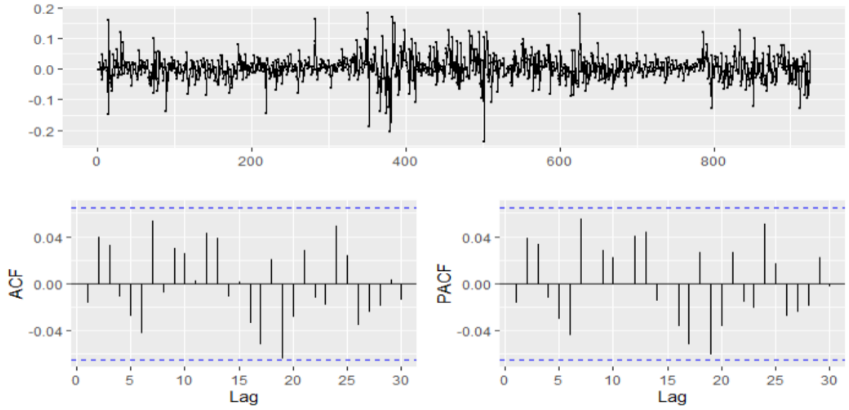
First, we change tesla closing price to be a time series object using the ts function in R. then we plot the ACF and the PACF.



Form this we can see that the time serise is not statinary and the variance is not constant as well.

Therefore, we will do difrencing and log transformation.

After doing the differencing and the log transformation we pot the time series again using ggtsdisplay function.

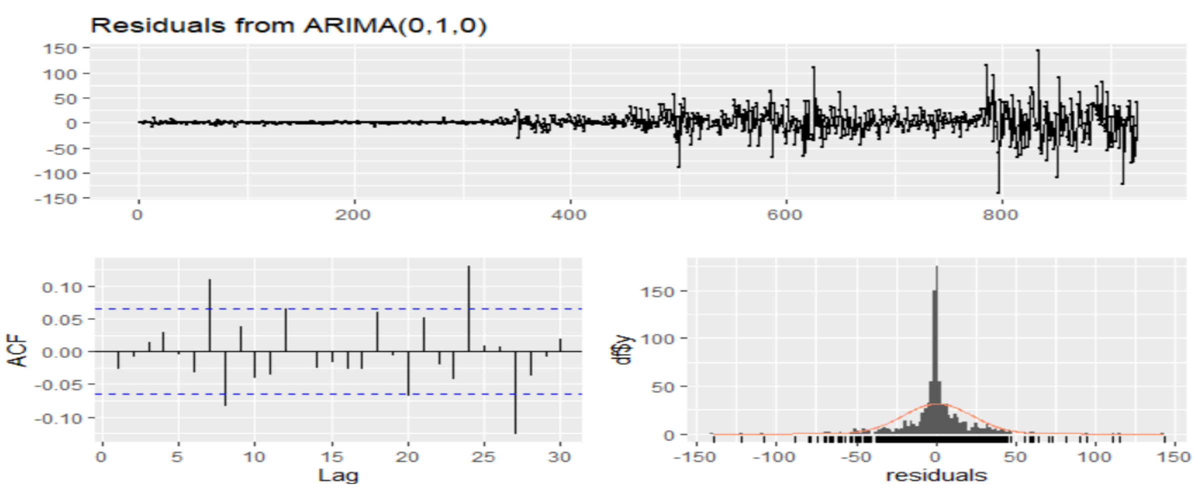


from this we can conclude that the time series is stationary, the variance is constant as well. Also, from the ACF and PACF, we see no significant spikes. This may be suggestive of a AR (0) term and MA (0).

Meaning that ARIMA (0,10) could be a good fit for our data. Moreover, using the auto Arima function suggest the same model as showing below.

```
Series: teslafinal$TSLA.Clos
ARIMA(0,1,0)

sigma^2 = 466.2:  log likelihood = -4149.94
AIC=8301.88   AICc=8301.89   BIC=8306.71
```

Then we plot the residuals form Arima (0,1,0).



the residuals statistic suggest that the suggested model is not all that good, since there is Autocorrelation between the error terms (Not white noise) and error terms is not normally distributed. Judging by the Ljung-Box test, we conclude that the p-value < 0.05 (significant) meaning that the model's residuals are autocorrelated. Put in other words, the Arima model suggested does have heteroscedasticity problem, hence we should do ARCH or GARCH model.

Fitting GARCH model.

Since some random days have very Hight volatility, GARCH will be the best model to explain it.

We need to calculate the annualized volatility and the rolling-window volatility of tesla closing price. This can be done either at the daily, monthly, quarterly frequency, etc.

starting with the standard GARCH model where we consider the conditional error term is a normal distribution. We use the function ugarchspec() for the model specification and ugarchfit() for the model fitting. For the standard GARCH model, we specify a constant to mean ARMA model, which means that arma0rder = c (0,0). We consider the GARCH (1,1) model and the distribution of the conditional error term is the normal distribution.

The following are the results of the estimation for the standard GARCH (1,1) model. With the normal distribution.

```
*---------------------------------*
*          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      0.002491    0.001248   1.9952 0.046025
omega   0.000098    0.000036   2.7234 0.006461
alpha1  0.080301    0.019674   4.0816 0.000045
beta1   0.865008    0.035228  24.5549 0.000000

Robust Standard Errors:
        Estimate  Std. Error  t value Pr(>|t|)
mu      0.002491    0.001381   1.8032 0.071351
omega   0.000098    0.000066   1.4852 0.137487
alpha1  0.080301    0.031765   2.5280 0.011472
beta1   0.865008    0.058795  14.7122 0.000000

LogLikelihood : 1665.861
```

The first table of the first part of the estimation (see table named "Optimal parameters") shows the optimal estimated parameters. This table shows the significance of the estimated parameter. It shows that the constant parameters tend to be significant, meaning that the constant parameters seem to be useful in this model setting. Also, this table show the loglikelihood, the bigger the better.

The second table presents the information criteria. It displays the Akaike (AIC), Bayes (BIC), Hannan-Quinn and Shibata criteria for the model estimation. The lower these values, the better the model is in terms of fitting.

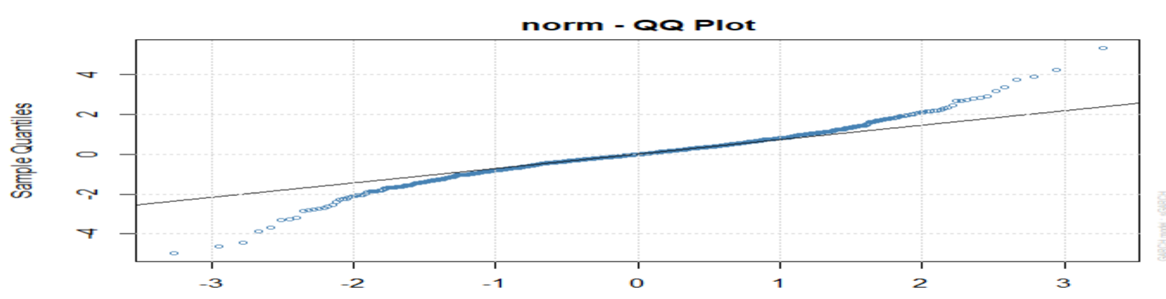
The next table presents the Ljung-Box test for testing the serial correlation of the error terms. The null hypothesis is that there is no serial correlation of the error terms. The decision rule is simple. Basically, if the p-value is lower than 5%, the null hypothesis is rejected. As we can see that the p-value is higher than 5%, meaning that there is not enough evidence to reject the null hypothesis. Then there is no serial correlation of the error term.

```
Weighted Ljung-Box Test on Standardized Residuals
------------------------------------
                        statistic p-value
Lag[1]                    0.01458  0.9039
Lag[2*(p+q)+(p+q)-1][2]   0.41932  0.7327
Lag[4*(p+q)+(p+q)-1][5]   1.14755  0.8255
d.o.f=0
H0 : No serial correlation

Weighted Ljung-Box Test on Standardized Squared Residuals
------------------------------------
                        statistic p-value
Lag[1]                     0.2137  0.6439
Lag[2*(p+q)+(p+q)-1][5]    2.1290  0.5883
Lag[4*(p+q)+(p+q)-1][9]    3.6708  0.6450
d.o.f=2
```

```
Adjusted Pearson Goodness-of-Fit Test:
------------------------------------
  group statistic p-value(g-1)
1    20     78.84    2.947e-09
2    30     89.40    4.502e-08
3    40    102.51    1.304e-07
4    50    113.09    5.600e-07
```

Another table that is interesting to check is the last table: "Adjusted Pearson Goodness of Fit", concerning the goodness of fit of the error. Indeed, it is useful to check if the error term follows the normal distribution. The null hypothesis is that the conditional error term follows a normal distribution. If the p-value is lower than 5%, the null hypothesis is rejected. As we can see, the normal distribution is by far rejected (as the p-value is close to zero.

we can see the QQ-plot and it show that the residuals are not that perfectly aligned with the straight line, meaning that the residuals do not follow the normal distribution.



To solve this issue, we will use GARCH (1,1) with the student's t-distribution.

Fitting GARCH model with Skewed student distribution:

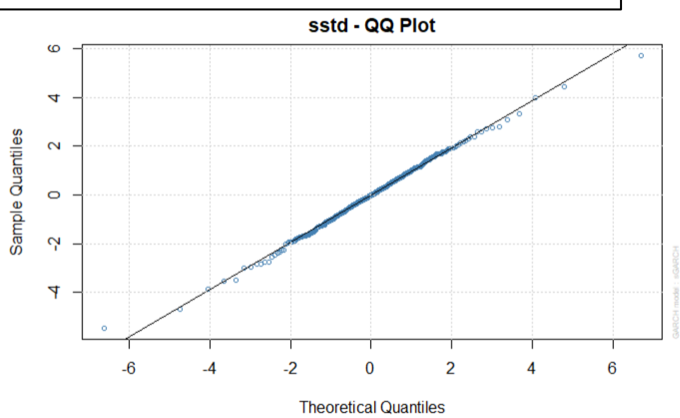
The following is the result for the GARCH (1,1) with the sstd distribution.

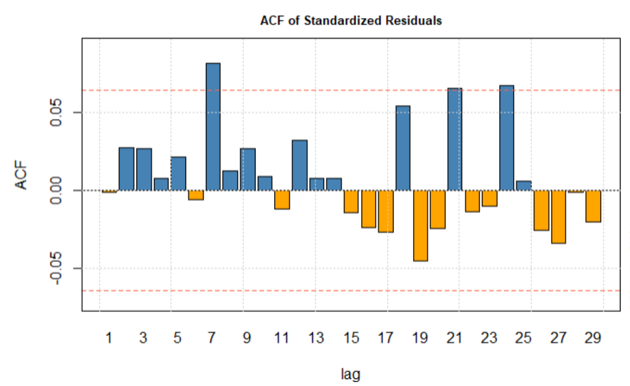
```
*              GARCH Model Fit                  *
*-----------------------------------------------*

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

Optimal Parameters
------------------------------------------------
        Estimate  Std. Error   t value Pr(>|t|)
mu      0.003131    0.001213    2.5821 0.009821
omega   0.000041    0.000036    1.1486 0.250703
alpha1  0.098791    0.041216    2.3969 0.016535
beta1   0.900191    0.044323   20.3097 0.000000
skew    1.004216    0.043534   23.0674 0.000000
shape   3.404494    0.442502    7.6937 0.000000

Robust Standard Errors:
        Estimate  Std. Error   t value Pr(>|t|)
mu      0.003131    0.001378   2.27162 0.023109
omega   0.000041    0.000075   0.54456 0.586059
alpha1  0.098791    0.076677   1.28840 0.197607
beta1   0.900191    0.094576   9.51819 0.000000
skew    1.004216    0.046606  21.54691 0.000000
shape   3.404494    0.447737   7.60377 0.000000

LogLikelihood : 1734.68
```

As we can see the loglikelihood is bigger than the one with the normal distribution. Meaning that this model is better than the previous one.

```
Adjusted Pearson Goodness-of-Fit Test:
------------------------------------
  group statistic p-value(g-1)
1    20     9.162       0.9707
2    30    21.041       0.8577
3    40    23.028       0.9803
4    50    36.959       0.8969
```

Now, we can see that on the last table, the p-values are higher than 0.05, meaning that the skewed student distribution is a good fit for the error term. Also, the AIC, BIC and Hannan-Quin value are lower than the one obtained from the previous setting (normal distribution case).





Now, we can see that the QQ-plot shows a more aligned distribution to the straight line and the return distribution follow the student distribution.

Fitting GARCH with regression:

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

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

Optimal Parameters
------------------------------------
        Estimate  Std. Error   t value Pr(>|t|)
mu      0.002386    0.000896  2.663580 0.007731
omega   0.000002    0.000006  0.321099 0.748135
alpha1  0.049995    0.001555 32.142132 0.000000
beta1   0.876345    0.008944 97.981570 0.000000
vxreg1  0.000000    0.000000  0.050984 0.959338
vxreg2  0.000000    0.000001  0.013008 0.989621
vxreg3  0.000000    0.000000  0.037334 0.970218
vxreg4  0.000000    0.000000  0.036079 0.971220
vxreg5  0.000000    0.000000  0.058082 0.953684
vxreg6  0.000000    0.000000  0.022939 0.981699
vxreg7  0.000000    0.000000  0.000000 1.000000
skew    1.000364    0.042730 23.411386 0.000000
shape   3.806729    0.210963 18.044571 0.000000
```

Then we fit GARCH with the regression model using the function (external. Regressors) trying to add some variable to the models thar might be influencing TESLA stock price. Our variables are DJI index, gold, oil, twitter, NIO (Chinese electric car company), lithium stock closing price and Tesla volume.

The following is the result for the model:

As we can see all the regression parameter are not-significant since they have a p-value bigger than 0.05. also, omega is not significant in this model.

```
LogLikelihood : 1668.722

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

Akaike       -3.5761
Bayes        -3.5083
Shibata      -3.5765
Hannan-Quinn -3.5502

Weighted Ljung-Box Test on Standardized Residuals
------------------------------------
                        statistic p-value
Lag[1]                     0.1785  0.6726
Lag[2*(p+q)+(p+q)-1][2]    0.8401  0.5531
Lag[4*(p+q)+(p+q)-1][5]    1.8748  0.6483
d.o.f=0
H0 : No serial correlation

Weighted Ljung-Box Test on Standardized Squared Residuals
------------------------------------
                        statistic   p-value
Lag[1]                      7.524 6.089e-03
Lag[2*(p+q)+(p+q)-1][5]    21.438 7.783e-06
Lag[4*(p+q)+(p+q)-1][9]    25.570 6.901e-06
d.o.f=2
```

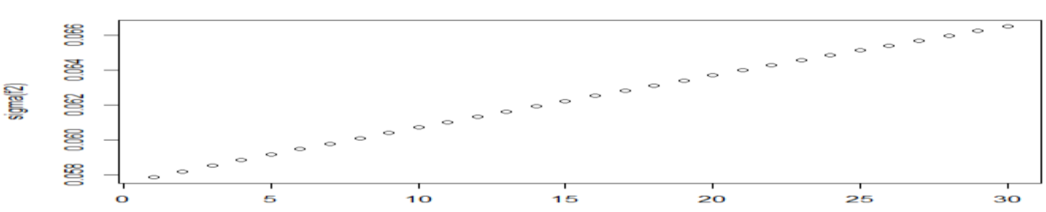
the loglikelihood for this model is smaller than the previous model (GARCH WITH NO REGRESSION). The AICc and the BIC are bigger than the other model. Hence, the other model is better than this one.

The Optimal GARCH model setting for the TESLA stock

After analyzing different models, we observed that the ARIMA (0,1,0) GARCH (1,1) with no regression., seems to work well for TESLA stock. Based on this model setting, we can see that all the parameters of the model are statistically significant. Indeed, their p-value is lower than 5. Also, the Akaike (AIC), Bayes (BIC), Hannan-Quinn and Shibata criteria are lower than the one observed from the other model setting. When testing the presence of serial correlation in the residuals, we can see that the p-value is greater than 5% for the different setting considered, meaning that there is no serial correlation in the residuals. Furthermore, the global test of the ARCH model shows that the ARCH model is globally significant as its global p-value is close to zero. For the goodness of fit of the residual to the considered skewed student distribution, we can see that the p-value is greater than 5%, meaning that there is not enough evidence to reject the fact that the residuals fit well that distribution.

Forecast

Now we fit our selected model to the data and run the forecast function.
When we run the forecast of the volatility for the next 30 days, we can observe that based on this model, we expect the volatility of TESLA to potentially keep increasing in the next 30 days as shows the graph below.



Running Multiple linear regression by itself:

When we run the multiple linear regression, we can see that some variables are significant in our model (twitter closing stock) Since their p-value is less than 0.05.

The following in the result.

```
Call:
lm(formula = train_data$TSLA.Close ~ ., data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.42597 -0.10730 -0.00611  0.10561  0.46888 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -20.03007    1.24075 -16.144   <2e-16 ***
DJI.Close    -0.30268    0.15169  -1.995   0.0463 *  
CL.Close     -1.75357    0.15027 -11.669   <2e-16 ***
GC.F.Close    3.73569    0.06855  54.494   <2e-16 ***
LIT.Close     1.41844    0.07037  20.156   <2e-16 ***
NIO.Close     0.22476    0.01811  12.410   <2e-16 ***
TWTR.Close   -0.05117    0.04773  -1.072   0.2840    
TSLA.Volume   0.15990    0.01207  13.242   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1552 on 792 degrees of freedom
Multiple R-squared:  0.9812,	Adjusted R-squared:  0.981 
F-statistic:  5891 on 7 and 792 DF,  p-value: < 2.2e-16

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  6.334   6.568   6.682   6.694   6.805   7.127 
```