AcF 633 - Python Programming for Data Analysis
Manh Pham

Group Project 2

7th March 2022 12pm to 21st March 2022 12pm (UK time)

This assignment contains one question worth 100 marks and constitutes 20% of the total marks for this course.

You are required to submit to Moodle a single .zip folder containing a single Jupyter Notebook .ipynb file OR a single Python script .py file, together with any supporting .csv files (e.g. input data files) AND a signed group coversheet. The name of this folder should be your group letter (e.g. GroupA.zip, where Group A is your group). Only one of the group members needs to upload the folder for your group.

Your submission .zip folder must be submitted electronically via Moodle on the **21st March 2022 12pm (UK time)**. Only one of the group members needs to submit the work for your group. Late submissions (without pre-approved extensions) will receive a zero mark.

Good Luck!

Question 1:

Refer back to the csv data file 'DowJones-Feb2022.csv' that lists the constituents of the Dow Jones Industrial Average (DJIA) index as of 9 February 2022 that was investigated in the group project 1. Import the data file into Python.

Using the order of your group letter in the alphabet (e.g. 1 for A, 2 for B, etc.) as a random seed, draw a random sample of 2 stocks (i.e. tickers) from the DJIA index excluding stock DOW.[1] Import daily Adjusted Close (Adj Close) prices for both stocks between 01/01/2008 and 31/12/2021 from Yahoo Finance. Compute the **log** daily returns (in %) for both stocks and drop days with NaN returns. Perform the following tasks.

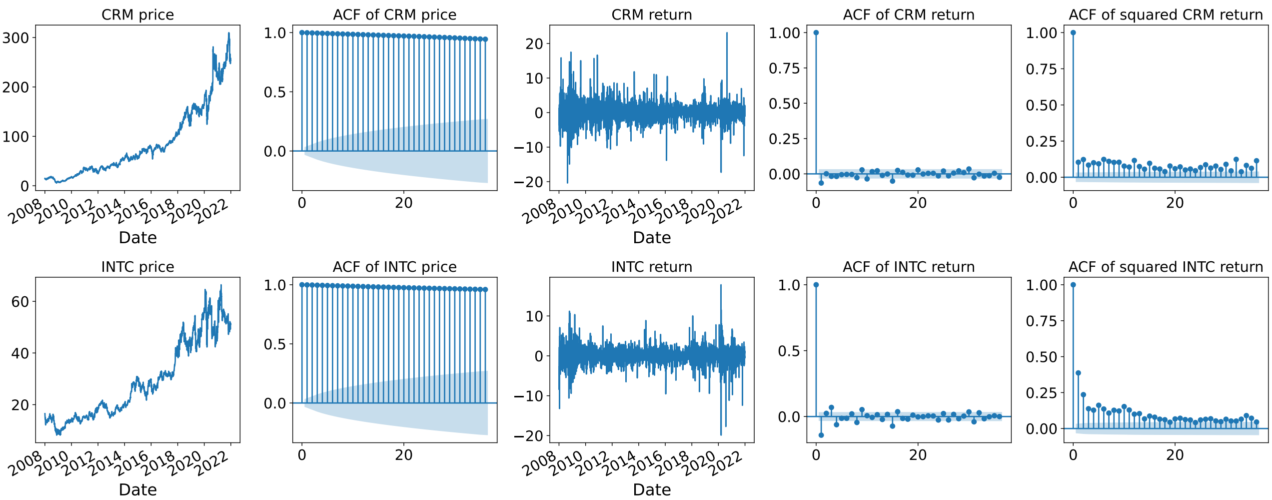
Task 1: Descriptive Analysis (Σ = **10 marks**)

1.1: Write code to plot a 2-by-5 subplot figure that includes:

Row 1: (i) Time series plot and (ii) ACF of the first stock price, (iii) time series plot and (iv) ACF of the first stock return, and (v) ACF of the square of first stock return.

Row 2: The same subplots for the second stock.

Your figure should look similar to the following for your sample of stocks. Comment on what you observe from the plots. (**6 marks**)



1.2: Using a 5% significance level, conduct an Augmented Dickey Fuller (adf) test (with regression = 'c') to check whether the price and return series of both stocks in your sample have a unit root (i.e. are integrated) or not, and if they are, determine their order of integratedness. Print a statement similar to the following for your stock sample. (**4 marks**)

```
CRM price is integrated of order 1.  CRM return is stationary.
INTC price is integrated of order 1.  INTC return is stationary.
```

[1] DOW only started trading on 20/03/2019.

Task 2: Return Modelling (Σ = **10 marks**)

2.1: Using data between 01/01/20008 and 31/12/2018 as in-sample data, write code to find the best-fitted ARMA(p, q) model for returns of each stock that minimizes AIC, with p and q no greater than 3. Print the best-fitted ARMA(p, q) output and a statement similar to the following for your stock sample.

```
Best-fitted ARMA model for CRM: ARMA(1,1) - AIC = 13114.7316
Best-fitted ARMA model for INTC: ARMA(2,2) - AIC = 11203.7960
```

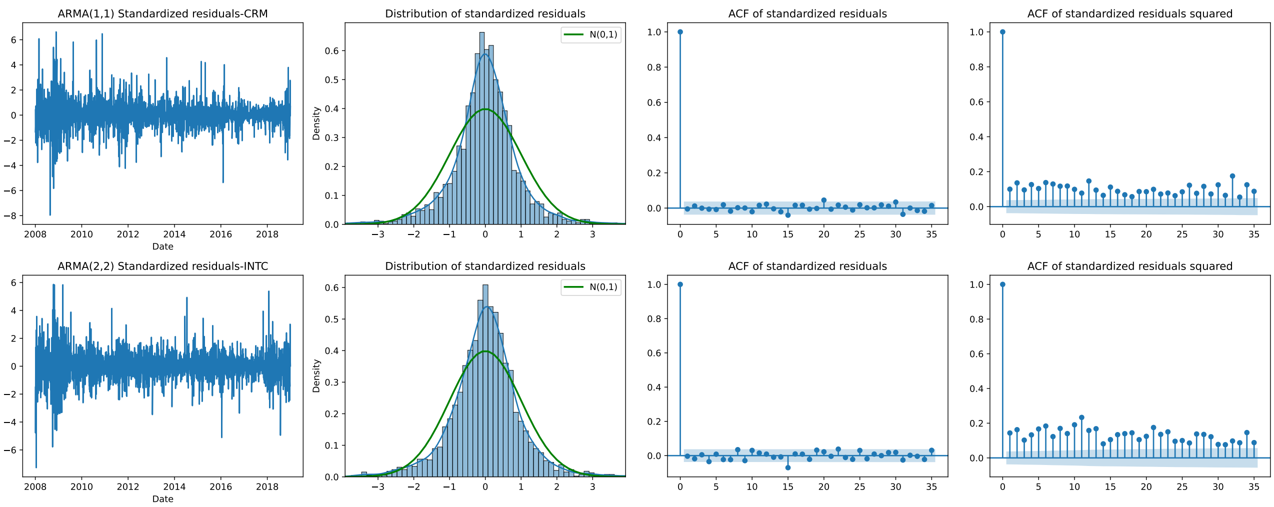
(**5 marks**)

2.2: Write code to plot a 2-by-4 subplot figure that includes the following diagnostics for the best-fitted ARMA model found in Task **2.1**:

Row 1: (i) Time series plot of the standardized residuals, (ii) histogram of the standardized residuals, fitted with a kernel density estimate and the density of a standard normal distribution, (iii) ACF of the standardized residuals, and (iv) ACF of the squared standardized residuals.

Row 2: The same subplots for the second stock.

Your figure should look similar to the following for your sample of stocks. Comment on what you observe from the plots. (**5 marks**)



Task 3: Volatility Modelling (Σ = **20 marks**)

3.1: Use the same in-sample data as in Task **2.1**, write code to find the best-fitted AR(p)-GARCH(p^*, q^*) model with Student's t errors for returns of each stock that minimizes AIC, where p is fixed at the AR lag order found in Task **2.1**, and p^* and q^* are no greater than 3. Print the best-fitted AR(p)-GARCH(p^*, q^*) output and a statement similar to the following for your stock sample.

```
Best-fitted AR(p)-GARCH(p*,q*) model for CRM: AR(1)-GARCH(1,1) - AIC
= 11989.4856
Best-fitted AR(p)-GARCH(p*,q*) model for INTC: AR(2)-GARCH(1,1) -
AIC = 10297.5848
```

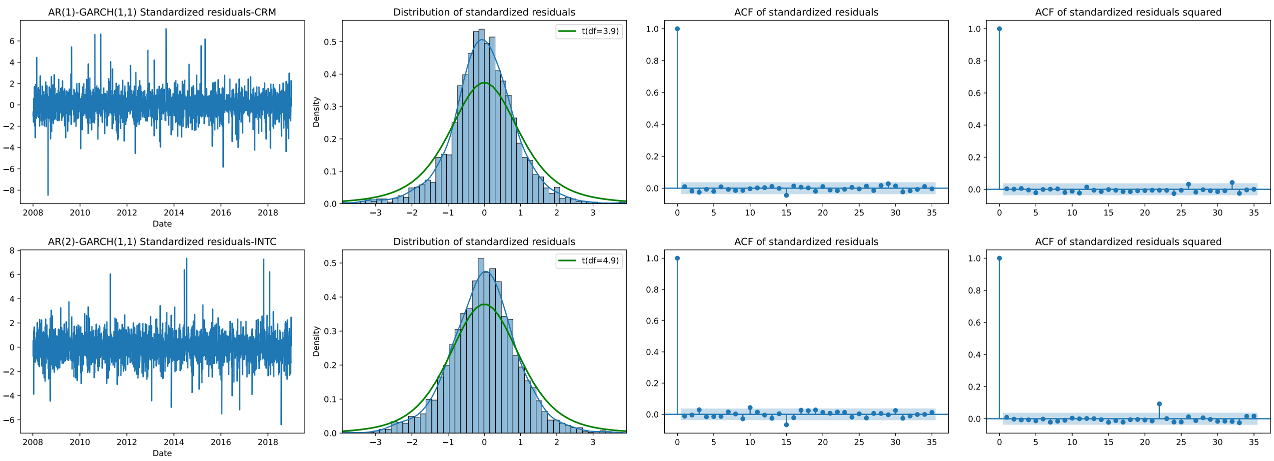
(**7 marks**)

3.2: Write code to plot a 2-by-4 subplot figure that includes the following diagnostics for the best-fitted AR-GARCH model found in Task **3.1**:

Row 1: (i) Time series plot of the standardized residuals, (ii) histogram of the standardized residuals, fitted with a kernel density estimate and the density of a standard normal distribution, (iii) ACF of the standardized residuals, and (iv) ACF of the squared standardized residuals.

Row 2: The same subplots for the second stock.

Your figure should look similar to the following for your sample of stocks. Comment on what you observe from the plots. (**5 marks**)

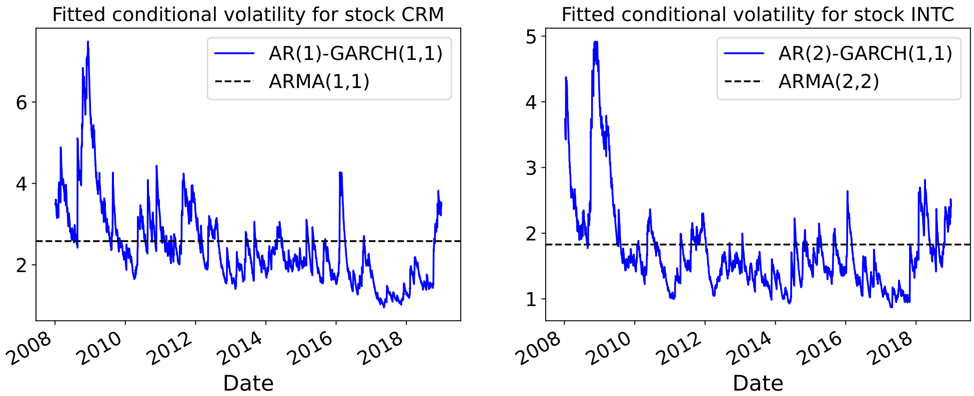


3.3: Conduct an ARCH LM test to check if there are any ARCH effects up to order 10 left in the standardized residuals of the best-fitted AR(p)-GARCH(p^*, q^*) model found in Task **3.2**. Using a 5% significance level, draw and print the conclusion similar to the following for your stock sample.

```
There are no ARCH effects up to order 10 left in the standardized
residuals of the AR(1)-GARCH(1,1) model for stock CRM.
There are no ARCH effects up to order 10 left in the standardized
residuals of the AR(2)-GARCH(1,1) model for stock INTC.
```

(**3 marks**)

3.4: Write code to plot a 1-by-2 subplot figure that shows the fitted conditional volatility implied by the best-fitted AR(p)-GARCH(p^*, q^*) model found in Task **3.1** against that implied by the best-fitted ARMA(p, q) model found in Task **2.1** for each stock in your sample. Your figure should look similar to the following. *(Hint: the conditional volatility implied by an ARMA(p, q) model is constant over time and equal to the estimate of the standard deviation (σ) of the error term.)*

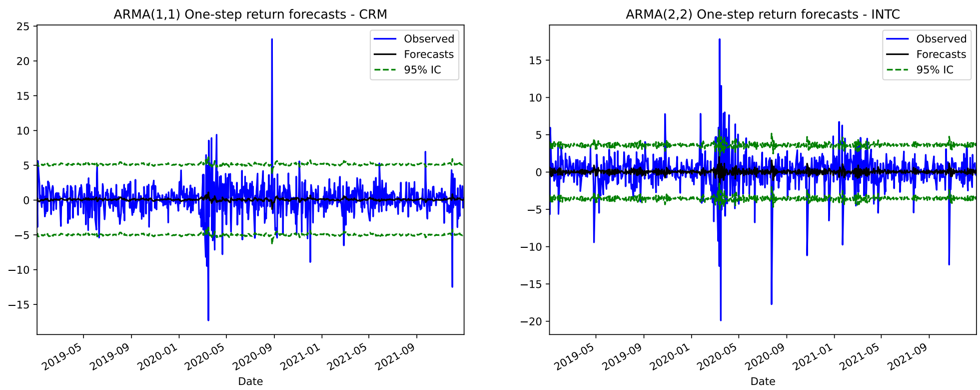


(**5 marks**)

Task 4: Return and Volatility Forecasting (Σ = **40 marks**)

4.1: Use data between 01/01/2019 and 31/12/2021 as out-of-sample data, write code to compute **one-step forecasts**, together with 95% confidence interval (CI), for the returns of each stock using the respective best-fitted ARMA(p, q) model found in Task **2.1**. You should extend the in-sample data by one observation each time it becomes available and apply the fitted ARMA(p, q) model to the extended sample to produce one-step forecasts. Do **not** refit the ARMA(p, q) model for each extending window.[2] For each stock, the forecast output is a data frame with 3 columns `f`, `fl` and `fu` corresponding to the one-step forecasts, 95% CI lower bounds, and 95% CI upper bounds. (**7 marks**)

4.2: Write code to plot a 1-by-2 subplot figure showing the one-step return forecasts found in Task **4.1** against the true values during the out-of-sample period for both stocks in your sample. Also show the 95% confidence interval of the return forecasts. Your figure should look similar to the following.

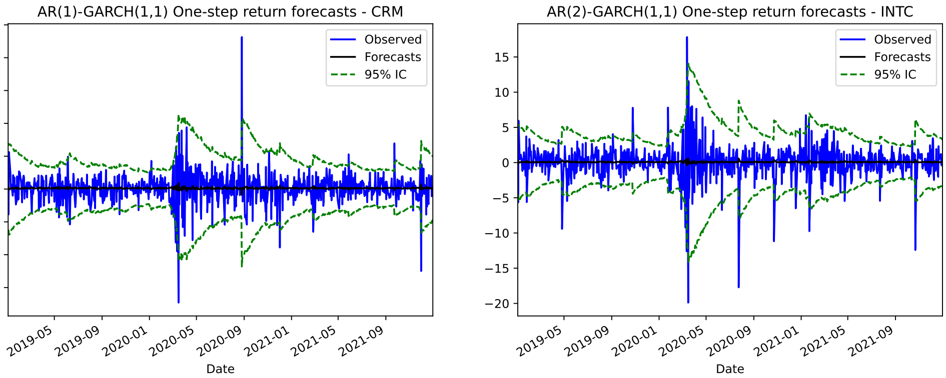


(**4 marks**)

4.3: Write code to produce **one-step analytic forecasts**, together with 95% confidence interval, for the returns of each stock using respective best-fitted AR(p)-GARCH(p^*, q^*) model found in Task **3.1**. For each stock, the forecast output is a data frame with 3 columns `f`, `fl` and `fu` corresponding to the one-step forecasts, 95% CI lower bounds, and 95% CI upper bounds. (**6 marks**)

[2] Refitting the model each time a new observation comes generally gives better forecasts. However, it slows down the program considerably so we do not pursue it here.

4.4: Write code to plot a 1-by-2 subplot figure showing the one-step return forecasts found in Task **4.3** against the true values during the out-of-sample period for both stocks in your sample. Also show the 95% confidence interval of the return forecasts. Your figure should look similar to the following.



(4 marks)

4.5: A return forecast is considered 'reasonably accurate' if the observed return falls within the 95% CI of the forecast. Compute the proportion of 'reasonably accurate' return forecasts implied by the best-fitted ARMA(p,q) and AR(p)-GARCH(p^*,q^*) models for each stock in your sample. Print a statement similar to the following for your stock sample.

```
Proportion of reasonably accurate forecasts for CRM: ARMA(1,1):
96.83%; AR(1)-GARCH(1,1):  95.90%
Proportion of reasonably accurate forecasts for INTC: ARMA(2,2):
92.47%; AR(2)-GARCH(1,1):  95.77%
```

(5 marks)

4.6: Denote by $e_{t+h|t} = y_{t+h} - \widehat{y}_{t+h|t}$ the h-step forecast error at time t, which is the difference between the observed value y_{t+h} and an h-step forecast $\widehat{y}_{t+h|t}$ produced by a forecast model. Four popular metrics to quantify the accuracy of the forecasts in an out-of-sample period with T' observations are:

1. Mean Absolute Error: MAE $= \dfrac{1}{T'} \sum_{t=1}^{T'} |e_{t+h|t}|$
2. Mean Square Error: MSE $= \dfrac{1}{T'} \sum_{t=1}^{T'} e^2_{t+h|t}$
3. Mean Absolute Percentage Error: MAPE $= \dfrac{1}{T'} \sum_{t=1}^{T'} |e_{t+h|t}/y_{t+h}|$
4. Mean Absolute Scaled Error: MASE $= \dfrac{1}{T'} \sum_{t=1}^{T'} \left| \dfrac{e_{t+h|t}}{\frac{1}{T'-1} \sum_{t=2}^{T'} |y_t - y_{t-1}|} \right|$.

The closer the above measures are to zero, the more accurate the forecasts. Now, write code to compute the four above forecast accuracy measures for one-step return forecasts produced by the best-fitted ARMA(p,q) and AR(p)-GARCH(p^*,q^*) models for each stock in your sample. For each stock, produce a data frame containing the forecast accuracy measures of a similar format to the following, with columns being the names of the above four accuracy measures and index being the names of the best-fitted ARMA and AR-GARCH model:

	MAE	MSE	MAPE	MASE
ARMA(1,1)	1.620	6.106	1.261	0.694
AR(1)-GARCH(1,1)	1.608	6.017	1.330	0.688

Print a statement similar to the following for your stock sample:

```
For CRM:
Measures that ARMA(1,1) model produces smaller than AR(1)-GARCH(1,1)
model:  MAPE
Measures that AR(1)-GARCH(1,1) model produces smaller than ARMA(1,1)
model:  MAE, MSE, MASE
```

(7 marks)

4.7: Using a 5% significance level, conduct the Diebold-Mariano test for each stock in your sample to test if the one-step return forecasts produced by the best-fitted ARMA(p,q) and AR(p)-GARCH(p^*,q^*) models are equally accurate based on the four accuracy measures in Task **4.6**. For each stock, produce a data frame containing the forecast accuracy measures of a similar format to the following:

	MAE	MSE	MAPE	MASE
ARMA(1,1)	1.620	6.106	1.261	0.694
AR(1)-GARCH(1,1)	1.608	6.017	1.330	0.688
DMm	2.962	2.360	-1.283	2.962
pvalue	0.003	0.019	0.200	0.003

where 'DMm' is the Harvey, Leybourne & Newbold (1997) modified Diebold-Mariano test statistic (defined in the lecture), and 'pvalue' is the p-value associated with the DMm statistic.

Print a statement similar to the following for your stock sample:

```
For CRM:
Model AR(1)-GARCH(1,1) produces significantly more accurate one-step
return forecasts than model ARMA(1,1) based on MAE.
Model AR(1)-GARCH(1,1) produces significantly more accurate one-step
return forecasts than model ARMA(1,1) based on MSE.
Models ARMA(1,1) and AR(1)-GARCH(1,1) produce equally accurate one-step
returns forecasts based on MAPE.
Model AR(1)-GARCH(1,1) produces significantly more accurate one-step
return forecasts than model ARMA(1,1) based on MASE.
```

(7 marks)

Task 5: (Σ = **20 marks**)

These marks will go to programs that are well structured, intuitive to use (i.e. provide sufficient comments for me to follow and are straightforward for me to run your code), generalisable (i.e. they can be applied to different sets of stocks (2 or more)) and elegant (i.e. code is neat and shows some degree of efficiency).