# City University of Hong Kong

## 2019/20 Semester B

## CS3481 Fundamentals of Data Science

## Project 2 – Report

## Group 7

| Name | Student ID |
|---|---|
| | |
| | |
| | |
| | |

# 1. Introduction

The world is full of uncertainty. An economy in an expansion cycle today might fall into recession tomorrow. An obvious example would be the 2007-08 financial crisis caused by deregulation of the financial industry and the famous Lehman Brothers' bankruptcy. The crisis was unexpected for most retail investors – individuals and small-volume trader. It is shown from statistics that, during the period from the stock market high on Oct.9.2007 to its bottom on March.9.2009, the two primary stock market indices – Dow Jones Industrial Average (DJI) and Standard & Poor's 500 (S&P 500) lost more than 50% of their value[1]. The huge loss in the indices imply, on average, a tremendous lost in investors' money. While reading the whole report, consider yourself a risk-averse investor. Given that different sectors/industries respond differently toward unexpected negative shocks, a natural tendency would be investing sectors that are less sensitive to extreme cases. In other words, select a sector which tends to be affected by crisis less than other industries. Therefore, it leads to the research on finding those specific sectors.

## 1.1 Problem description

In this report, we aimed to find a sector that was less volatile compared to another in 2007-08 financial crisis period. Due to time and complexity constraints, we mainly focused on two sectors – financial services and healthcare. The problem is divided into following steps:

- Build time-series models on daily sector stock index for pre-crisis period
- Select the model that best describe the data based on various measures
- Predict the sector stock index price during crisis using the model selected
- Evaluate the divergence between real index price and predicted price

The conclusion will be based on the evaluation, which greater divergence greater volatility and vice versa. The smaller divergent sector stock index during the crisis period will be the outcome of the research. Rationale behind is that smaller the difference means actual prices during the negative shock period is less different from the predicted prices – where the prediction model was built based on data prior to the crisis period, then the unexpected impact/impact the model cannot capture of the shock on the specific index price is less. Therefore, the sector with smaller divergence is considered less volatile relatively to other sector; which, is the goal of our research. The whole study will be built based on Python.

## 1.2 Data Source

We use Dow Jones U.S. Financial Services Index (^DJUSGF) daily close price and Dow Jones U.S. Health Care Index (^DJUSHC) daily close price as our source data. The reason for choosing Dow Jones sector index is because of its representativeness in capturing sector stocks' price behaviour. The timeframe of the chosen data is from 2002-08-01 to 2008-03-04, which covers the period before and in the financial crisis. All data are extracted from Bloomberg Terminal. Moreover, we download research data from Fama-French library to elaborate our models.

## 1.3 Proposed Model

[1] Gold, H. (2018, September 19). This is the biggest lesson investors should learn from the 2008 financial crisis. Retrieved April 8, 2020, from https://www.marketwatch.com/story/this-is-the-biggest-lesson-investors-should-learn-from-the-2008-financial-crisis-2018-09-18

We proposed three models to fit the daily close price time series and do forecasting. The models are 1) Capital Asset Pricing Model (CAPM) with Generalized Auto Regressive Conditional Heteroskedasticity (GARCH); 2) Fama-French three-factor model (FF) with GARCH; and 3) Auto Regressive Integrated Moving Average (ARIMA). The details about each model will be discussed in the next section.

# 2. Modelling

## 2.1 Data Pre-processing

The sample is divided into three groups – training, testing and forecasting, with the ratio roughly be 9:1:1. The specification of each set is shown as below.

| | Time period | | # of instances |
|---|---|---|---|
| Training set | Pre-crisis | 2002-08-01 to 2007-01-30 | 1174 |
| Test set | Pre-crisis | 2007-01-31 to 2007-09-04 | 155 |
| Forecast set | During the crisis | 2007-09-05 to 2008-03-04 | 130 |

Table 1: Data split in different time period

The set division for each sector index dataset is the same. The forecast set is used for making the final judgment about the model prediction ability during the crisis.

In the following part, we will do the pre-processing on the training set.

### 2.1.1 Data Cleaning

We deleted the unnecessary columns and rename column names and replace the index with dates. To deal with the missing value, we found only the first day has missing data thus we simply delete the first row. Because the data we deal with is time series data, we only have numerical data. The statistic table of our current dataset is below.

| | DJUSGF | DJUSHC |
|---|---|---|
| Count | 2089 | 2089 |
| Mean | 590.273887 | 296.872647 |
| Std | 166.383904 | 33.644827 |
| Min | 194.45 | 204.71 |
| 25% | 475.92 | 271.16 |
| 50% | 600.42 | 298.64 |
| 75% | 798.5 | 320.51 |
| Max | 924.1 | 367.73 |

Table 2: Detailed Information on our dataset

### 2.1.2 Data Visualization

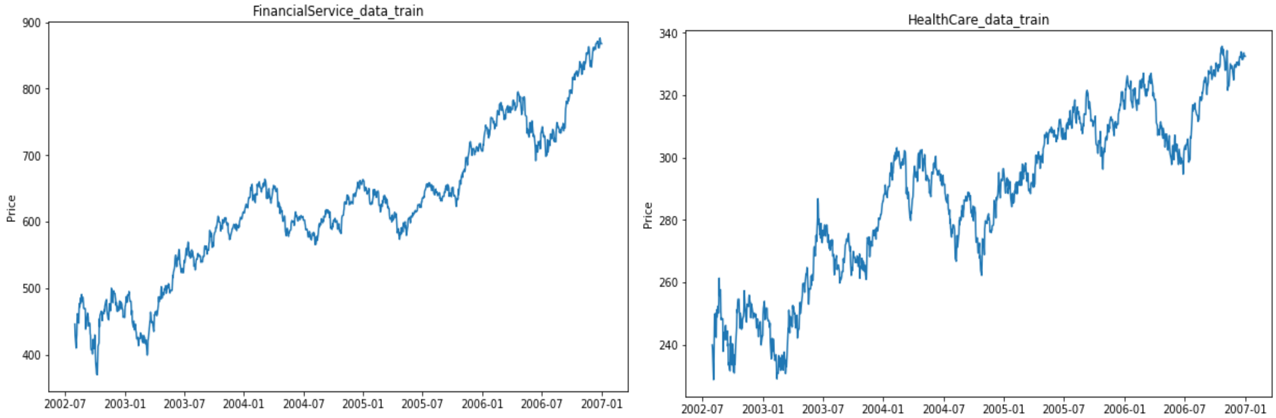


Figure 1: Price trend of two sector indexes. Both indexes are in an up-trend in pre-crisis period.

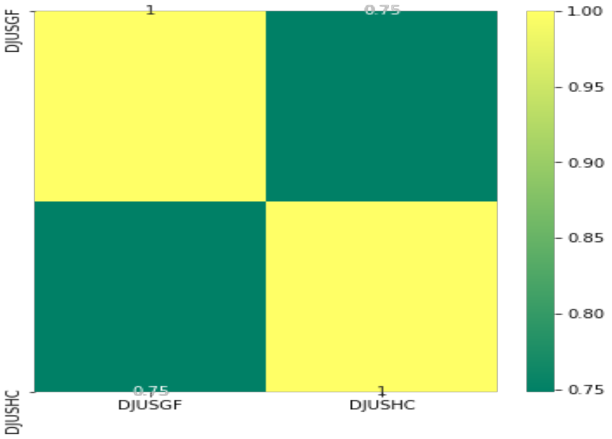


Figure 2: Correlation between two sector indexes

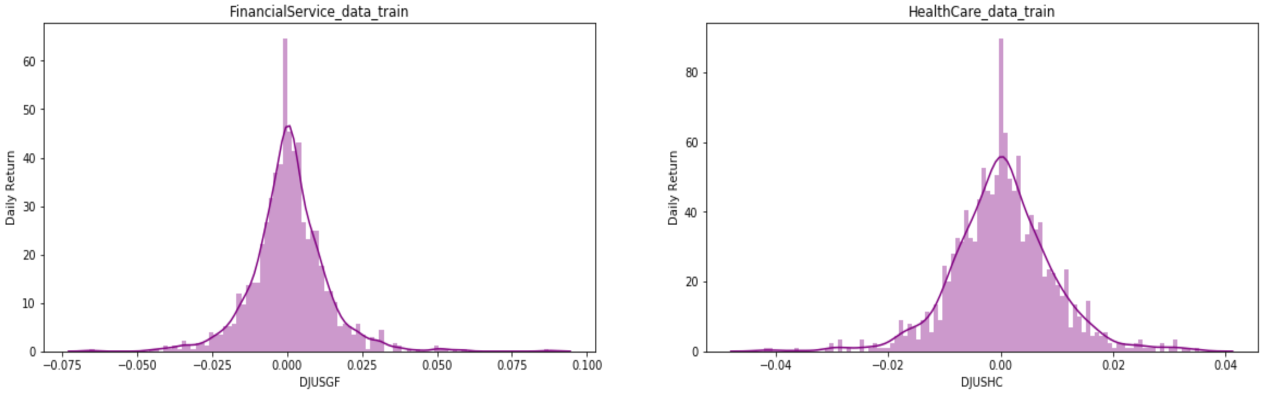


Figure 3: Daily return of two sector indexes

From the above figure, we can see that the return of the two indexes are roughly follow a standard normal distribution.

### 2.1.3 Stationary Test

Stationarity need to be ensured before applying any statistical model on our dataset. There are two primary way to determine whether the time series is stationary:

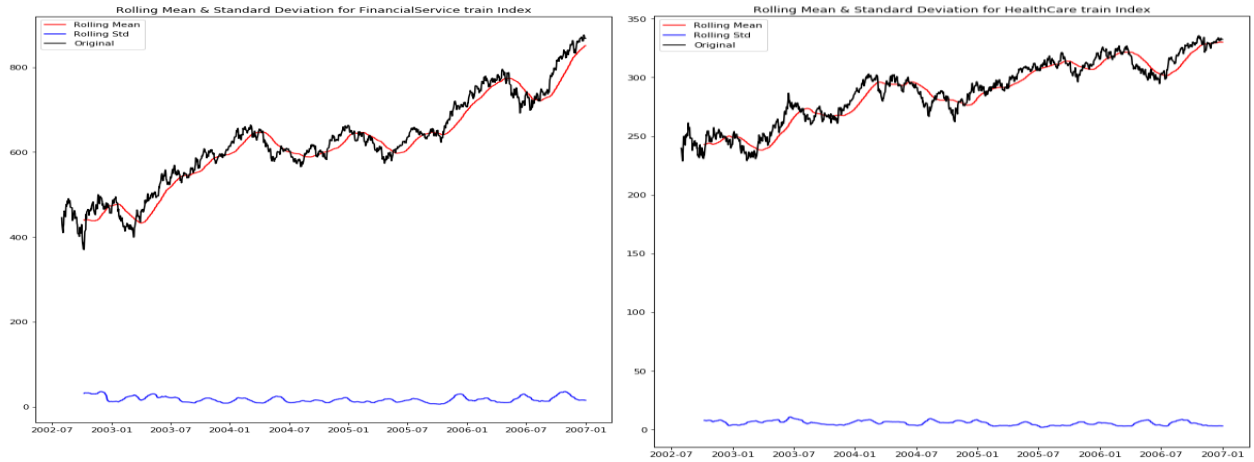
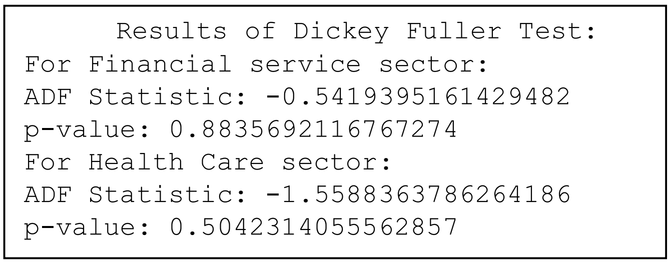


Figure 4: Rolling Statistics: By looking at the rolling mean and rolling standard deviation plot to see if they remain constant with the time (stationary)

The rolling mean increase with time rolling but the standard deviation is fairly constant with time. Therefore, we can conclude that the time series is not stationary.

Augmented Dickey–Fuller test: The Null hypothesis of this test is the time series is non-stationary. The test gives us a p-value. If the P-value is less than the critical values, we can reject the null hypothesis and say that the series is stationary.



```
       Results of Dickey Fuller Test:
For Financial service sector:
ADF Statistic: -0.5419395161429482
p-value: 0.8835692116767274
For Health Care sector:
ADF Statistic: -1.5588363786264186
p-value: 0.5042314055562857
```

Figure 5: Results of Dickey Fuller Test

As the p-value is greater than the common threshold (0.05). Thus, at 5% significant level, we can conclude that the time series is not stationary.

### 2.1.4 Data Transformation

To achieve stationarity, we have two measures to transform data.

i. Log Scale Transformation: Taking the log of the dependent variable and then subtract the moving average to remove trend.
The formula is S'= Log(s)-moving average.

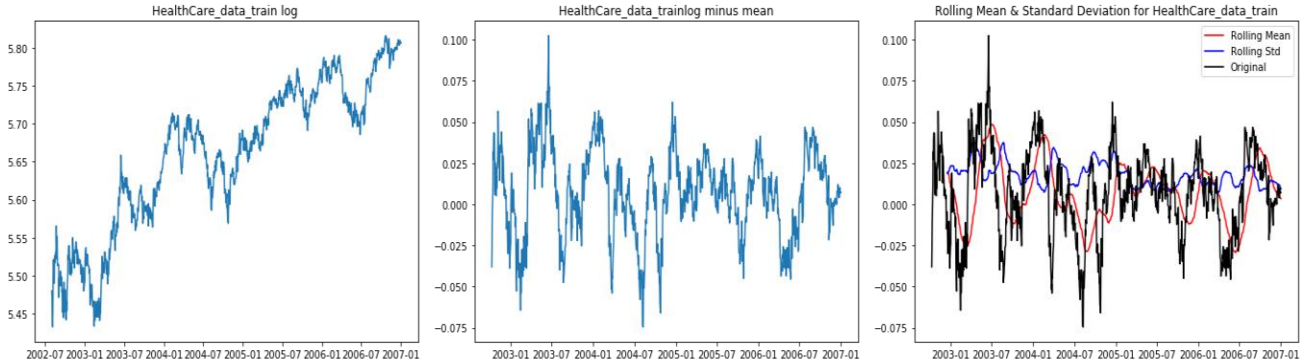
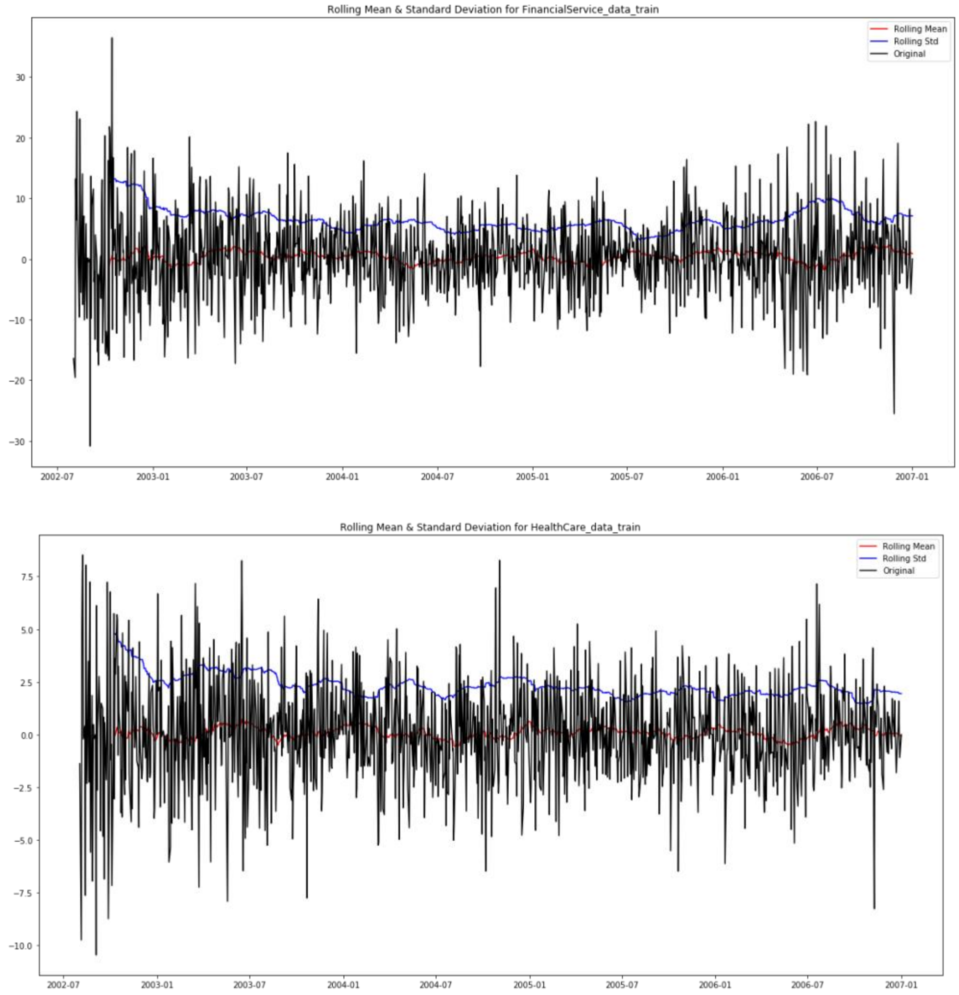


Figure 6: Log scale transformed data

From the charts above, and according to the ADF test (p-values are small than 0.05), we have sufficient evidences to say the time series is stationary.

ii. Time Shift Transformation: Take difference between current price and lagged one price. E.g. null, (x1−x0), (x2−x1), (x3−x2), (x4−x3), ..., (xn−xn−1)



Rolling Mean & Standard Deviation for FinancialService_data_train

Rolling Mean & Standard Deviation for HealthCare_data_train

Figure 7: Time shift transformed data

From the charts above, and according to the ADF test (p-values are small than 0.05), we have enough evidence to say the time series is stationary.

As we can see, additional work is needed for F-F model and CAPM. Because of the discrepancies of time-series index between independent variables and dependent variables, we generate data-frame based on the common index independent variables shared with dependent variables. The omitted data only have slight impact on our parameter's estimation, so we are fine. To obtain daily return from stock price, one should take the first order differencing and neglect first row as it's null variable. Then divided the differencing sequence by the stock price yesterday. Formula is below.

$$R_i(t) = \frac{[S(t) - S(t-1)]}{S(t-1)}$$

## 2.2Model

### 2.2.1 Model 1 : Capital Asset Pricing Model (CAPM) with Generalized Auto Regressive Conditional Heteroskedasticity (GARCH)

***Model introduction***

The model composes two models – 1) CAPM and 2) GARCH, which 1) is modelling the mean process and 2) is evaluating the volatility evolution.

**CAPM** was developed by econometrician William Sharpe, Jack Treynor, John Lintner and Jan Mossin in the early 1960s[2]. The model answered the primary question in finance – the relationship between risk of an investment and corresponding expected return. The risks in finance genre are usually divided into two kinds, which are systematic risk and unsystematic risk. On the one hand, systematic risk is associated with the whole market, which cannot be eliminated. On the other hand, unsystematic risk is related to a specific industry or security, which can be eliminated by diversifying the portfolio. In other words, holding stocks belong to different sectors may help lower the unsystematic risk.

The fundamental idea behind the model is that fully diversified portfolio theoretically has zero unsystematic risk. Intuitively, not all kinds of risk are related to the return. For example, if your portfolio currently consists two stock –Johnson & Johnson (JNJ) and Visa Inc A (V). Without the idea behind CAPM, when calculating the expected return of the portfolio, you will consider all individual risk of each stock. However, there is a possibility that part of the risk exposed by holding JNJ is offset by holding V. Therefore, the risk being offset should not be considered when evaluating return. Hence, CAPM models the expected return by taking systematic risk into account.

The formula for CAPM is given as below.

$$E(R_i) = R_f + \beta_i\big(E(R_m) - R_f\big)$$

Where
$R_i(t)$: simple return of a stock i at time t;
$R_m(t)$: simple return of market portfolio at time t;
$E(R_i)$: expected return of the asset i;
$E(R_m)$: expected return of the market;
$R_f$: risk-free rate, which usually considered to be treasury bond interest;
$\beta_i$: volatility of an asset or a portfolio of assets compare to the market volatility, $\frac{Cov(R_i,R_m)}{Var(R_m)}$;

Since we apply the model in one asset at a particular time t, the restated model is

$$E\big(R_i(t)\big) - R_f = \beta_i\Big(E\big(R_m(t)\big) - R_f\Big)$$

**GARCH** model was introduced by Bollerslev[3] in 1986. The model is commonly used for modelling financial time-series data with focus on the volatility process and exhibit the volatility clustering effect. For better understanding of the model, we will explain it word by word. Generalized Auto Regressive Conditional Heteroskedasticity: first, Auto Regressive means that the model is a regression model with both dependent variable and independent variable itself. Typically, the difference between the two variables is the

[2] Perold, A. F. (2004). The Capital Asset Pricing Model. *Journal of Economic Perspectives*, *18*(3), 3–24. Retrieved from https://pubs.aeaweb.org/doi/pdf/10.1257/0895330042162340

[3] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics, 31*(3), 307-327. Retrieved from https://www.sciencedirect.com.ezproxy.cityu.edu.hk/science/article/pii/0304407686900631

time. For example, the relationship between Apple stock's volatility at time t and its variance at time t-1 is a kind of autoregressive process.

Then, for heteroskedasticity, it is an opposite concept with homoscedasticity. While in statistics, homoscedasticity means the error term in the linear regression model has constant variance[4], heteroskedasticity means a nonconstant variance structure for the error term.[5] Conditional heteroskedasticity applies when volatility in the future periods is unable to be identified. Therefore, ARCH is an autoregressive model with unidentifiable changing variance in the error term.

In terms of the generalization, it assumes the shifting variance of the innovation follows an autoregressive moving average (ARMA) model. In short, ARMA model means volatility in time t is associated with volatility in time t-p, where p stands for the lagged time and p<t. Hence, GARCH models the volatility evolution and captures the volatility clustering effect.

The formula for GARCH (p, q) is given as below.

$$\{X_t = \sigma_t \varepsilon_t, \qquad \varepsilon_t \sim N(0,1) \qquad\qquad \sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \beta_i \sigma_{t-i}^2 + \sum_{j=1}^{p} \alpha_j X_{t-j}^2$$

Where
$X_t$: a time series variable;
$\sigma_t^2$: conditional volatility at time t;
q: number of lagged conditional error terms to be considered in modeling volatility in time t;
p: number of lagged time-series variable to be considered in modeling $\sigma_t$.
$\alpha_0, \beta_i, \alpha_j$: Parameters to be estimated in GARCH (p, q) model

**Doob–Meyer Decomposition Theorem** is well known that a discrete parameter super-martingale (down-trending stochastic process) is the sum of a martingale (random walk) and a decreasing and time-deterministic process. Meyer gave necessary and sufficient conditions under which a continuous parameter super-martingale is the sum of a martingale and a natural decreasing process. Similarly, sub-martingale can be expressed in terms of the summation of a martingale and an increasing general process. Therefore, decomposition of stock return into mean-process and volatility-process is vital when evaluating stock return.[6]

## ***Methodology***

### Part A: Mean process modelling

To model the stock index price, we need to figure out the distribution of the price curve, therefore, the most crucial part is to model the mean and variance. In this model, we use CAPM as the mean model stated in previous part. Apply ordinary least square method to the regression model, we can figure out estimated mean process $\{\{\widehat{E(R_\iota(t))}\}_{t=1}^{T}$.

[4] Homoscedasticity. (2013). In D. Rutherford, *Routledge Dictionary of Economics* (3rd ed.). Routledge. Credo Reference:http://ezproxy.cityu.edu.hk/login?url=https://search.credoreference.com/content/entry/routsobk/homoscedasticity/0?institutionId=6601

[5] Chan, N. (2011). Heteroskedasticity. In *Wiley Series in Probability and Statistics* (pp. 105-122). Hoboken, NJ, USA: John Wiley & Sons.

[6] Meyer, Paul-André (1963). "Decomposition of Supermartingales: the Uniqueness Theorem". Illinois Journal of Mathematics. 7 (1): 1–17.

**Part B: Volatility process modelling**

We assume the difference between the real return and the expected return is an unknown $\varepsilon_t$,

$$R_i(t) - E[R_i(t)] = \varepsilon_t$$

To model the unknown residual time series $\{\varepsilon_t\}_{t=1}^{T}$, we apply GARCH model,

$$\{\varepsilon_t = \sigma_t Z_t, \qquad Z_t \sim N(0,1) \qquad \sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \beta_i \sigma_{t-i}^2 + \sum_{j=1}^{p} \alpha_j \varepsilon_{t-j}^2$$

Notice that $\varepsilon_t$ is actually the de-mean process of stock return, Doob-decomposition implies the stochastic term of stock return comes from the residuals. Therefore, we can apply GARCH model to capture the innovation of volatility. To determine automatically the order of GARCH, we can make use of *"pdmarima"* package in Python.

If we apply GARCH (p, q) to the sequences of residuals, we obtain $\{|\sigma_t|\}_{t=1}^{T}$. Hence, to give signs to estimated shock, we estimate $\varepsilon_t$ as $\varepsilon_t = |\sigma_t| * X(t)$, where $X(t)$=1 if $R_i(t) - E[R_i(t)]$ >0; and $X(t) = -1$ otherwise.

Since the value of $\sigma_t^2$ is deterministic at time $t-1$, the rationale of using $|\sigma_t| * X(t)$ to estimate residual lays in:

$$Var(I_{t-1}) = Var(I_{t-1}) = \sigma_t^2 * Var(I_{t-1}) = h_t$$

$$E[(I_{t-1})] = E[|\sigma_t|] * E[(I_{t-1})] = \sigma_t^2 * E[X(t)] = \sigma_t^2 * X(t)$$

The reason $E[X(t)] = X(t)$ is that we already have the data set for actual return and expected return from the first F-F regression model. From the illustration, we know residual in regression model has the same conditional mean and variance as our estimator.

In general, our regression model has a form of:

$$\widehat{R_i(t)} - r_f = \beta_i * (R_m(t) - r_f) + \widehat{\varepsilon_t}$$

**Part C: Forecasting**

Since we use regression method to model time-series data and we assume we won't get any information after Lehman's bankruptcy, to make forecast, ARIMA technique in adapted in this case to forecast mean-process. For volatility-process, we will still use GARCH model to make prediction. For detail of ARIMA forecasting, the detail will be explained in ARIMA model.

Notes: our estimation for sequence of residuals $\{\varepsilon_t\}_{t=1}^{T}$ is $\left\{\widehat{\varepsilon_t} = \sqrt{h_t} * X(t)\right\}_{t=1}^{T}$, one could improve the estimation of volatility by modeling sequence $\{\varepsilon_t - \widehat{\varepsilon_t}\}_{t=1}^{T}$. In the view of balance between goodness of fit and complexity, we will suspend in this layer of differencing.

Finally, we can model sequences of stock price $\{S(t)\}_{t=1}^{T}$ in a way such that:

$\left\{\widehat{S(t')} = S(0) * \prod_{t=1}^{t'} \widehat{R_\iota(t)}\right\}_{t=1}^{T}$, where S(0) is the initial stock price.

**Part D: Potential Limitation**

Mean-process: It's hard to determine expected value of stock return simply by one factor.

Volatility-process: By Ito's Lemma[7] and well-known assumption of Black-Scholes model, the volatility of a stock is increasing with time-elapsing, in other words:

$$\frac{dS}{S} = \mu dt + \sigma dB$$

Where:
$\frac{dS}{S}$: Stock return over a small-time interval with length $dt$.
$\mu$: Mean process of stock, it can be a time-deterministic function.
$dB$: Change of a random variable with zero mean and 1 standard deviation.
It implies that 95% confidence interval will be wider as time passing. However, as we can see in Model evaluation, we obtain a relatively stable confidence interval.

### 2.2.2 Model 2 : Fama-French Three Factor Model (F-F) with Generalized Auto Regressive Conditional Heteroskedasticity

***Model introduction***

The model composes two models – 1) F-F model and 2) GARCH, which 1) is modelling the mean process and 2) is evaluating the volatility evolution. This model is an extension of Model 1 and adapt similar technique to model mean process and volatility process.

An empirical study by Fama and French (1992) shows that covariance of portfolio returns and market return does not explain changes on portfolio excess returns. They find that covariance has little or no power in terms of explaining cross-sectional variations in equity returns. [8]

Fama and French (1993) argue that anomalies relating to the CAPM are captured by the three-factor model. They observed two classes of stocks have tended to perform better than the market portfolio: (i) small market capitalization and (ii) stocks with a high book-to-market ratio. Thenceforth, Fama and French proposed an extension format for capital asset pricing model (CAPM) by adding two more factors SMB and HML into well-known CAPM equation.[9]

***Methodology***

**Part A: Mean process modelling**

To capture mean-process of stock return, Fama-French's model (F-F) states that:

$$E[R_i(t)] - r_f = a_0 + \beta_1 * \left(R_m(t) - r_f\right) + \beta_2 * SMB_t + \beta_3 * HML_t$$

Where:
$R_i(t), r_f, R_m(t)$ are the same factors in CAPM.
$SMB_t$ : Difference of return between small cap and big cap companies.
$HML_t$ : Difference of return between companies with high book-to-market ratio and those with low book-to-market ratio.
$a_0, \beta_1, \beta_2, \beta_3$ are parameters to be estimated.

**Part B: Volatility process modelling**

[7] Itō, K. (1944): Stochastic integral. Proc. Imp. Acad. Tokyo 20, 519-524.

[8] Fama, E. F.; French, K. R. (1992). "The Cross-Section of Expected Stock Returns". The Journal of Finance.

[9] Fama, E. F.; French, K. R. (1993). "Common risk factors in the returns on stocks and bonds". Journal of Financial Economics. 33: 3–56

Similar to model 1, to model the unknown shock $\varepsilon_t$, we can apply GARCH (p, q) to model the sequences of residuals $\{\varepsilon_t\}_{t=1}^{T}$,

As GARCH model equations are in mathematical form, in python, we can do it by changing the formula a little bit:

$$\widehat{R_i(t)} - r_f = a_0 + \beta_1 * (R_m(t) - r_f) + \beta_2 * SMB_t + \beta_3 * HML_t + \sqrt{h_t} * X(t)$$

$$X(t) = \begin{cases} 1, & R_i(t) - E[R_i(t)] \geq 0 \\ -1, & R_i(t) - E[R_i(t)] < 0 \end{cases}$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \beta_i \sigma_{t-i}^2 + \sum_{j=1}^{p} \alpha_j \varepsilon_{t-j}^2$$

**Part C: Forecasting**

The relationships that a regression model estimates might be valid for only the specific population that we sampled. As we assume that we have all information up to 2007.9.15, the observation of variables in regression become unavailable, which makes it impossible to predict future trend. To make solid forecasting, we will adapt the technique in ARIMA section. But rather to sequences $\{S(t)\}_{t=1}^{T}$, estimated $\{R_i(t)\}_{t=1}^{T}$ will be our target as we believe investor's positive and negative feelings do creep into the stock market and have an effect on stock market performance. As a result, the stock price deviates from its normal pattern, hence estimated $\{R_i(t)\}_{t=1}^{T}$ is a rational choice to make forecasting.

**Part D: Limitation**

Although F-F model has a better forecasting on mean-process, it still has the volatility drawback mentioned in CAPM model. Strong evidences show that simple return of stock follows geometric Brownian motion, it implies future price of financial stocks has a lognormal probability distribution and its volatility (standard deviation) is positively related to time elapsing. [10]In other words, the length of 95% confidence interval will gets wider. However, the volatility of GARCH is a mean-reverting process which means it converges to a constant number in long run.

### 2.2.3 Model 3: Autoregressive Integrated Moving Average Model

***Model introduction***

Before introducing the ARIMA model, we have three general models frequently used in economic industry.

| | | |
|---|---|---|
| Autoregressive Model AR(p) | **X**$_t$ = **a**$_0$ + $\sum_{j=1}^{p} a$**jXt − j** +$\epsilon_t$. | Assumes the value of the dependent variable on the current day depends on the past values. |
| Moving Average Model MA(q) | **X**$_t$ = **a**$_0$ + $\sum_{i=1}^{q}$ **aj**$\epsilon_{t-i}$ +$\epsilon_t$. | Assumes the value of the dependent variable on the current day depends on the previous days error terms |
| Autoregressive Moving Average Model ARMA(p, q) | **X**$_t$ = **a**$_0$ + $\sum_{j=1}^{p} a_j$**X**$_{t-j}$+ $\sum_{i=1}^{q} b$ $_{j}\epsilon_{t-I}$ +$\epsilon_t$. | It is a combination of above two models. |

Table 3: Information on three models used in economic industry

Where { $\epsilon_t$ }~White noise N(0, $\delta$2).

[10] Øksendal, Bernt K. (2002), Stochastic Differential Equations: An Introduction with Applications, Springer, p. 326

Autoregressive Integrated Moving Average Model (ARIMA) is a model adding differencing to an ARMA model. In this case, we have already done the first-order differencing which has transformed the time series into stationary ones.
Three integers (p, d, q) are typically used to parametrize ARIMA models.
p: number of autoregressive terms (AR order)
d: number of non seasonal differences (differencing order)
q: number of moving-average terms (MA order)
d has already been chosen to be 1.

***Methodology***

For determining p and q, We use Partial Auto Correlation Function (PACF) and Auto Correlation Function (ACF) separately.

ACF express the correlation between the observations at the current point in time and the observations at all previous points in time. It helps choose the optimal order of MA model(q).

PACF expresses the correlation between observations made at two points in time while accounting for any influence from other data points. It helps choose the optimal order of AR model(p).

After the operation of these steps, we got the final model for the two sectors' indices.

**Financial Services Index：**

ARIMA(8,1,8)

$R_t=0.3656+0.4581R_{t-1}+0.1899R_{t-2}+0.3282R_{t-3}-0.6151R_{t-4}+0.2878R_{t-5}+0.1601R_{t-6}+0.4792R_{t-7}-0.9692R_{t-8}-0.4655\varepsilon_{t-1}-0.1804\varepsilon_{t-2}-0.3219\varepsilon_{t-3}+0.5980\varepsilon_{t-4}-0.3143\varepsilon_{t-5}-0.1634\varepsilon_{t-6}-0.4741\varepsilon_{t-7}+0.9775\varepsilon_{t-8}+\varepsilon_t$
where $\{\varepsilon t\} \sim WN(0,\sigma^2)$

**Health Care Index：**

ARIMA(8,1,3)

$R_t=0.0813-1.1111R_{t-1}-1.0967R_{t-2}-0.8274R_{t-3}-0.2221R_{t-4}-0.1281R_{t-5}-0.0486R_{t-6}-0.0384R_{t-7}+0.0359R_{t-8}+1.0587\varepsilon_{t-1}+0.9491\varepsilon_{t-2}+0.6156\varepsilon_{t-3}+\varepsilon_t$ where $\{\varepsilon t\} \sim WN(0,\sigma^2)$

# 3 Model selection

Since we have already trained our three models (ARIMA Model, Fama French three factor Model and Capital Asset Pricing Model) using our training data of financial services and health care which is from 2002-08-01 to 2007-01-30, then we wanted to select one or more effective models to predict the tendency of these two stock indices during the crisis.

In this section, we observed MAPE(Mean Absolute Percentage Error) and AIC(Akaike Information Criterion) to estimate the error, and compared the effects among these three models based on our test data which is from 2007-01-31 to 2007-09-04.

- **MAPE**

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n} \left|\frac{\hat{y}_i - y_i}{y_i}\right|$$

From the above equation, we can see that the range of MAPE is $[0, +\infty)$. We want the value to be small, and if the value is 0%, it can be regarded as a perfect model actually.

- **AIC**

AIC is an estimator of out-of-sample prediction error, which can evaluate the quality of each model relative to each of the other model, if given scads of models for the dataset. AIC evaluates the relative amount of information lost by a given model: the less information a model loses, the higher the quality of that model. AIC encourages the goodness of data fitting, but it tries to avoid overfitting. Hence, this method is a good way for model selection. Here is the AIC value of the model:

$$AIC = 2k - 2ln\,(\hat{L})$$

k is the number of estimated parameters in the model and $\hat{L}$ is the maximum value of the likelihood function for the model. Usually, we choose the method whose AIC value is relatively small.

For our three models, we calculated the both AIC values and MAPE values. Using the error function, we determined that the level for MAPE below 5% is optimal and we could accept the model if the AIC value does not exceed the minimum 150.

| | AIC value | MAPE |
|---|---|---|
| FF for Financial Service | 1186 | 4.66% |
| FF for Health Care | 886 | 3.87% |
| CAPM for Financial Service | 1256 | 4.79% |
| CAPM for Health Care | 997 | 4.02% |
| ARIMA for Financial Service | 1134 | 4.48% |
| ARIMA for Health Care | 778 | 2.39% |

Table 4: AIC value and MAPE of three models for the two indices

The table illustrates that ARIMA model has the lowest MAPE and AIC value for either financial service or health care compared with the other two models. However, the gap of MAPE and AIC value between these models is tiny and reasonable, and MAPE values are all below 5%, which means all of these models perform well. Meanwhile, the disparity between the largest AIC value and the smallest is no larger than 150 for both two indices. Therefore, it is rational that we can take the three models into consideration and make our prediction for the two indices during the period of financial crisis.

## 4 Prediction

During our prediction, we only consider the financial crisis happened from 2007-09-05 to 2008-03-04 using the three models which we have tested and selected in the above section.

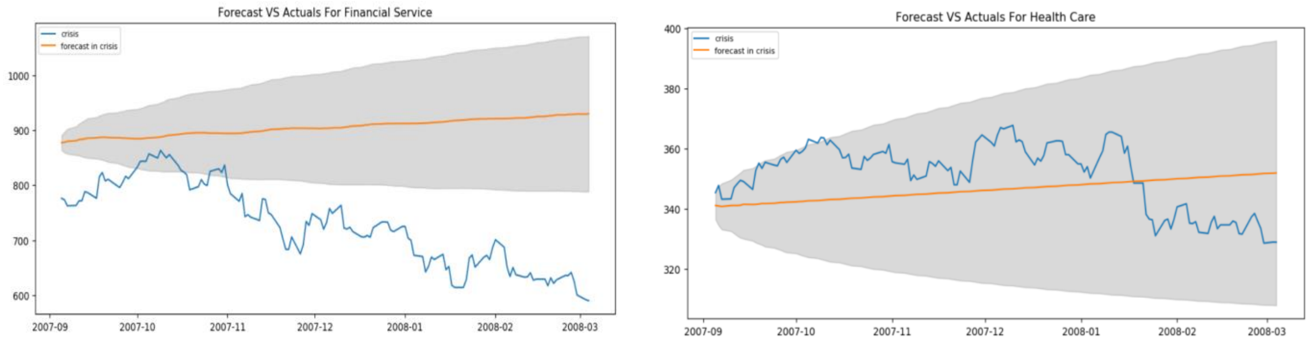


Figure 8: Forecast VS Actual for Financial Service (left)

and Health Care(right) using ARIMA

ARIMA model is a linear regression. In the figures, the gray region represents the 95% confidence interval. The blue and orange part are the tendency of two indices suffering the crisis and predications respectively. It is obviously that the financial service has a dramatic decline and is more sensitive compared with the health care during the crisis.

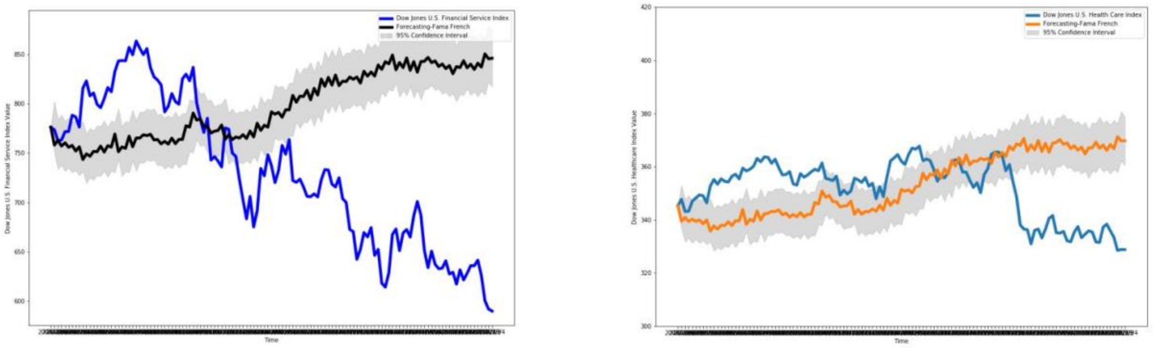


Figure 9: Forecast VS Actual for Financial (left) and Health Care(right) using Fama French

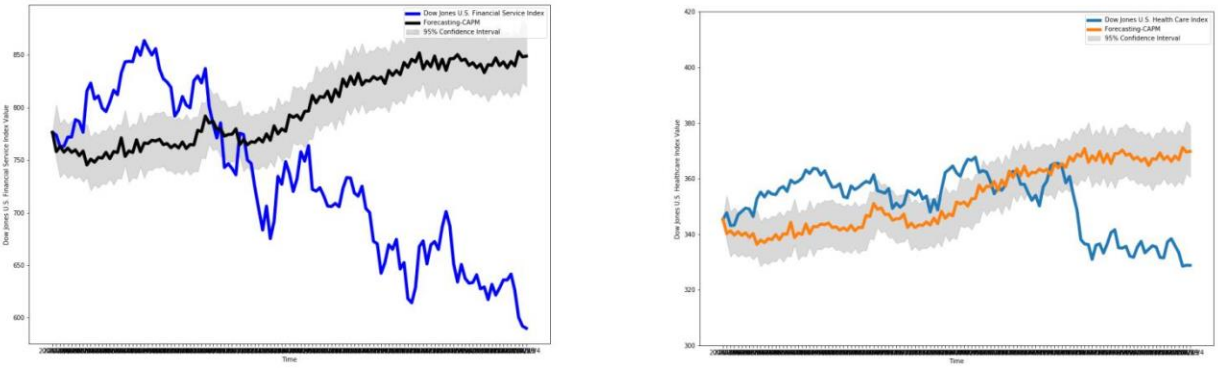


Figure 10: Forecast VS Actual for Financial Service(left) and Health Care(right) using CAPM

The patterns of CAPM and Fama French model are similar for predicting the two indices during the crisis. Considering the time-varying volatility, we could make our predictions better. Like ARIMA model, the gray region is also used to represent the 95% confidence interval. From the above figures, we can easily see that the prediction has a larger gap for Financial Service than Health Care either using CAPM or Fama French model.

# 5 Evaluation

In order to represent the gaps more accurately and intuitively, we introduced two mathematical means to measure them in the period of crisis. By this, we can gain a deep insight into the two indices during the period of crisis.

| | Chi-square measure: FS | Chi-square measure: HC | Euler Square: FS | Euler Square: HC |
|---|---|---|---|---|
| CAPM | 660.14 | 33.49 | 1403.23 | 216.91 |
| ARIMA | 1589.26 | 17.46 | 2244.35 | 156.09 |
| FF | 644.72 | 33.88 | 1385.69 | 218.15 |
| Average | 964.70 | 28.28 | 1677.76 | 197.05 |

Table 5: Chi-square and Euler square measure for the three models

The table compares the chi-square distance and Euler distance of the two indices for the three models. We can apparently see that the values of both chi-square distance and euler distance of financial service are much larger than those of health care based on the three models and the average values. In other words, the gap between financial service and normal prediction is much larger compared with the situation of health care. It means that financial service is more sensible and tends to be affected more by the crisis. However, the financial crisis makes a relatively slight difference to the health care, although the trend is also dropping overall.

# 6 Further discussion

Our study has adopted three models to analyse the problem we stated in the very beginning, however, there are some limitations.

1) The scope of our study is relatively narrow. As we only focus on two sectors – financial services and health care, the research cannot give broader insights on other sectors.

2) Lack of variety in model selection. All three models we examined are based on autoregressive-type, which limit the possibility of other better-fitted models.

3) The date chosen as the starting point of the financial crisis is arbitrary. As the date for which the crisis in 2007-08 started is not a fixed objective date, we chose the date based on various online research and our subjective opinion. This could cause divergence in the model performance.

To cope with the problems, we suggest introducing the Long Short-Term Memory (LSTM) model to predict six major sector indices price during financial crisis period. Qiu et al. (2020) mentioned that LSTM neural networks are suitable for financial time series (e.g. stock price). As stock price series are nonstationary and nonlinear, neural network in deep learning is becoming popular in predicting the series. Due to time and complexity constraints, we did not include this model in our study. However, it could be a direction for future research.

# 7 References:

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics, 31*(3), 307-327. Retrieved from https://www.sciencedirect.com.ezproxy.cityu.edu.hk/science/article/pii/0304407686900631

Chan, N. (2011). Heteroskedasticity. In *Wiley Series in Probability and Statistics* (pp. 105-122). Hoboken, NJ, USA: John Wiley & Sons.

Fama, E. F.; French, K. R. (1992). "The Cross-Section of Expected Stock Returns". The Journal of Finance.

Fama, E. F.; French, K. R. (1993). "Common risk factors in the returns on stocks and bonds". Journal of Financial Economics. 33: 3–56

Gold, H. (2018, September 19). This is the biggest lesson investors should learn from the 2008 financial crisis. Retrieved April 8, 2020, from https://www.marketwatch.com/story/this-is-the-biggest-lesson-investors-should-learn-from-the-2008-financial-crisis-2018-09-18

Homoscedasticity. (2013). In D. Rutherford, *Routledge Dictionary of Economics* (3rd ed.). Routledge. Credo Reference: http://ezproxy.cityu.edu.hk/login?url=https://search.credoreference.com/content/entry/routsobk/homoscedasticity/0?institutionId=6601

Itō, K. (1944): Stochastic integral. Proc. Imp. Acad. Tokyo 20, 519-524.

Meyer, Paul-André (1963). "Decomposition of Supermartingales: the Uniqueness Theorem". Illinois Journal of Mathematics. 7 (1): 1–17.

Øksendal, Bernt K. (2002), Stochastic Differential Equations: An Introduction with Applications, Springer, p. 326

Perold, A. F. (2004). The Capital Asset Pricing Model. *Journal of Economic Perspectives*, *18*(3), 3–24. Retrieved from https://pubs.aeaweb.org/doi/pdf/10.1257/0895330042162340