Statistical Analysis of 2 Financial Indexes

——An Analysis of Dow Jones Industrial Average (DJIA) and S&P 500

Team 3

Part 1 Preview of Dow Jones Industrial Average and S&P 500

In this part we analyze two data sets of Dow Jones Industrial Average (DJIA) index and S&P 500 index from **Yahoo! Finance.**

Q1. DJIA index value is equal to the sum of component equity prices divided by devisor.

Symbol	Company Name	Last Price
WMT	Walmart Inc.	119.04
UNH	UnitedHealth Group Incorporated	244.91
HD	The Home Depot, Inc.	234.38
XOM	Exxon Mobil Corporation	69.25
UTX	United Technologies Corporation	142.96
VZ	Verizon Communications Inc.	60.37
MRK	Merck & Co., Inc.	82.26
DIS	The Walt Disney Company	130.9
MSFT	Microsoft Corporation	140.73
NKE	NIKE, Inc.	90.92
JNJ	Johnson & Johnson	128.35
MCD	McDonald's Corporation	194.61
TRV	The Travelers Companies, Inc.	130.43
JPM	JPMorgan Chase & Co.	126.03
CVX	Chevron Corporation	118.67
V	Visa Inc.	177.85
IBM	International Business Machines Corporation	135.44
PFE	Pfizer Inc.	36.77
CSCO	Cisco Systems, Inc.	46.9
AAPL	Apple Inc.	246.58
PG	The Procter & Gamble Company	123.25
BA	The Boeing Company	339.83
GS	The Goldman Sachs Group, Inc.	214.23
KO	The Coca-Cola Company	53.75
AXP	American Express Company	118.26
WBA	Walgreens Boots Alliance, Inc.	55.42
DOW	Dow Inc.	50.48
MMM	3M Company	166.09
CAT	Caterpillar Inc.	139.73
INTC	Intel Corporation	56.46
Sum of price		3974.85
Devisor		0.147445684
Calculated DJIA = Sum of price/Devisor		26958.06
Official DJIA		26958.06

Table1: Components of DJI, price as of 2019-10-25 market close

Conclusions:

- The index today consists of 30 U.S. blue-chip stocks. The name "Industrial" is largely historical, as most stocks in this index are not from manufacturing industries, but rather from all the major sectors except utilities and transportation.
- Due to lagged refreshing frequency, calculating index by one's own may provide advantage in trading.

Q2. S&P 500 index introduction.

S&P 500 Index Introduction

The S&P 500 or Standard & Poor's 500 Index is a market-capitalization-weighted index of the 500 largest U.S. publicly traded companies. The index is widely regarded as the best gauge of large-cap U.S. equities. And S&P 500 includes some of the components of Dow Jones 30.

Weighting Formula and Calculation for the S&P 500

The S&P 500 uses a market capitalization weighting method, giving a higher percentage allocation to companies with the largest market capitalizations.

$$Company\ Weighting\ in\ S\&P = \frac{Company\ market\ cap}{Total\ of\ all\ market\ caps}$$

Determination of the weighting of each component of the S&P 500 begins with summing the total market cap for the index.

1. Calculate the total market cap for the index by adding all the market caps of the individual companies.
2. The weighting of each company in the index is calculated by taking the company's market capitalization and dividing it by the total market cap of the index.
3. For review, the market capitalization of a company is calculated by taking the current stock price and multiplying it by the company's outstanding shares.
4. Fortunately, the total market cap for the S&P as well as the market caps of individual companies is published frequently on financial websites saving investors the need to calculate them.

S&P 500 Index Construction

The market capitalization of a company is calculated by taking the current stock price and multiplying it by the outstanding shares. The S&P only uses free-floating shares, meaning the shares that the public can trade. The S&P adjusts each company's market cap to compensate for new share issues or company mergers. The value of the index is calculated by totaling the adjusted market caps of each company and dividing the result by a divisor. Unfortunately, the divisor is proprietary information of the S&P and is not released to the public.

However, we can calculate a company's weighting in the index, which can provide investors with valuable information. If a stock rises or falls, we can get a sense as to whether it might have an impact on the overall index. For example, a company with a 10% weighting will have a greater impact on the value of the index than a company with a 2% weighting.

The Widely Quoted S&P 500

The S&P 500 is one of the most widely quoted American indexes because it represents the largest publicly traded corporations in the U.S. The S&P 500 focuses on the U.S. market's large-cap sector

and is also a float-weighted index, meaning company market capitalizations are adjusted by the number of shares available for public trading.

Limitations of the S&P 500 Index

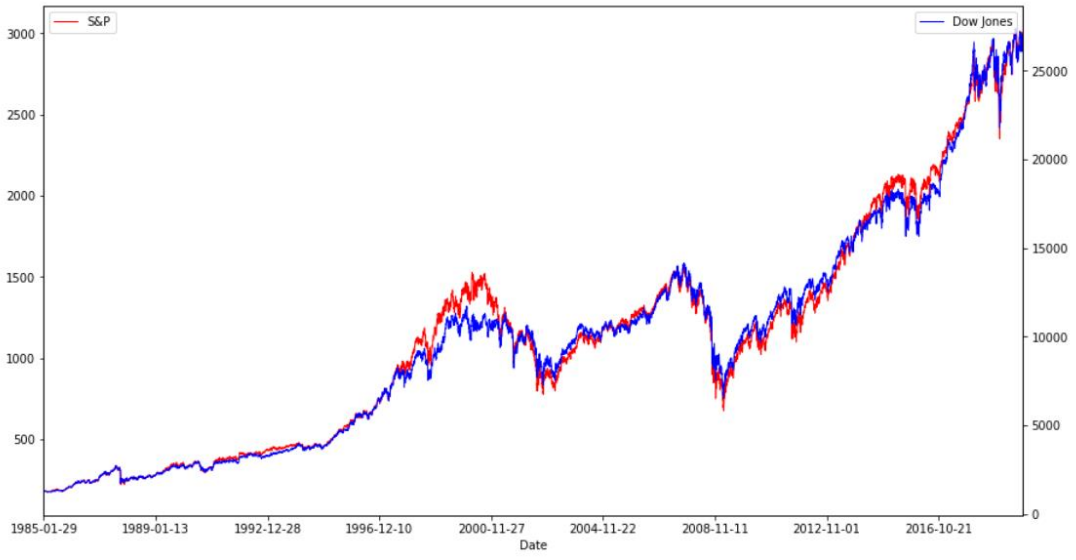
One of the limitations to the S&P and other indexes that are market-cap weighted arises when stocks in the index become overvalued meaning they rise higher than their fundamentals warrant. If a stock has a heavy weighting in the index while being overvalued, the stock typically inflates the overall value or price of the index.

A rising market cap of a company isn't necessarily indicative of a company's fundamentals, but rather it reflects the stock's increase in value relative to shares outstanding. As a result, equal-weighted indexes have become increasingly popular whereby each company's stock price movements have an equal impact on the index.

KEY TAKEAWAYS:

1. The DJIA is 30 U.S. stocks picked by the S&P Dow Jones Indices.
2. The S&P 500 is 500 U.S. stocks picked by an S&P Dow Jones Indices board.
3. The DJIA is calculated through a method of simple mathematical averages.
4. The S&P 500 is calculated by giving weights to each stock according to their market value.
5. While both indexes are used by investors to determine the general trend of the U.S. stock market, the S&P 500 is more encompassing, as it includes a greater sample of total U.S. stocks.

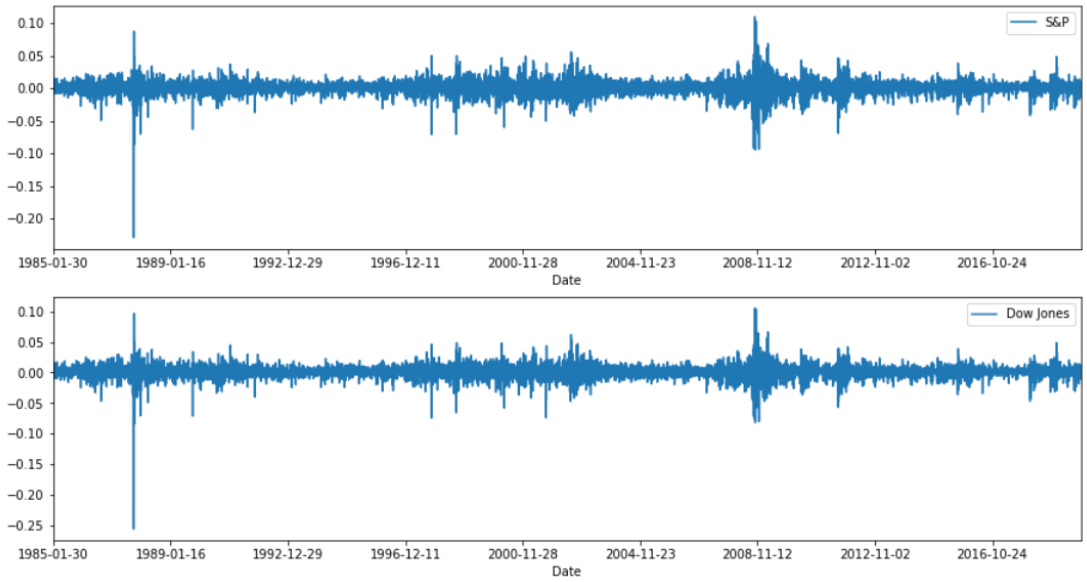
Q3. Time series of the two indexes



Conclusions:

1. S&P moves highly consistent with the Dow Jones index in the long run.
2. During the major inconsistencies in late 1990s and 2015-2016, S&P 500 reflects the general market better as a result of greater sample of total U.S. stocks.

Q4. Time series of log returns



Q5. Sample mean and variance of log return

	S&P	Dow Jones
$\hat{\mu}$	0.032%	0.035%
$\hat{S}^2$	$1.259\%^2$	$1.203\%^2$

Table2: Mean and variance of log return of index

Q6. Annualized average and volatility of log return

	S&P	Dow Jones
annualized log return	8.12%	8.73%
annualized volatility	17.81%	17.41%

Table3: Annualized log return feature of index

Q7. Sample skewness and sample kurtosis

	S&P	Dow Jones
skewness	-1.3	-1.7
kurtosis	31.0	44.6

Table4: Skewness and Kurtosis of 2 index

Conclusions:

1. The negative skewness for both indexes shows a left skewness, indicating a longer tail in left.
2. The high kurtosis shows a feature of fat tail, which predicts movements of three or more standard deviations more frequently than a normal distribution. It's in violation of assumptions in many traditional strategies of asset pricing.

Q8. Jarque-Bera test statistic (JB) and test of normality at the 5% significance level.

To test whether the kurtosis and skewness of the equation match a normal distribution, we introduced Jarque-Bera test[1]:

Decision rule: Reject H0 when test statistic lower or greater than the critical value. Otherwise, do not reject H0.

$$H_0 : JB = 0 \text{ (skewness being zero and the excess kurtosis being zero)}$$

$$H_1 : JB \neq 0 \text{ (At least skewness or kurtosis 6= 0)}$$

$$\alpha = 0.05$$

Test statistic:

$$JB_{S\&P500} = 287349$$

$$JB_{Dow\ Jones} = 634213$$

Critical Value:

$$X^2_{(k-1)(r-1)} = X^2_{(2-1)(8756-1)} = 31.41$$

Conclusion: We reject H_0 as the test statistic > critical value. It indicates that the residual distribution is not normal.

---

[1] Jarque–Bera test is a goodness-of-fit test of whether sample data have the skewness and kurtosis matching a normal distribution. The test statistic JB is defined as:

$$\mathrm{JB} = \frac{\mathrm{n-k}}{6}\left(\mathrm{S}^2 + \frac{1}{4}(\mathrm{K}-3)^2\right),$$

where S is the sample skewness, K is the sample kurtosis.

If the data comes from a normal distribution, the JB statistic asymptotically has a chi-squared distribution with two degrees of freedom, so the statistic can be used to test the hypothesis that the data are from a normal distribution. The null hypothesis is a joint hypothesis of the skewness being zero and the excess kurtosis being zero.

Part 2 Log return relationship between DJIA and S&P 500

This part analyzes the relationship between the log returns of DJIA and S&P 500 indexes. In particular, the two-sample t-test is a statistical algorithm to determine if two population means are equal. An application is to test if the log return of S&P 500 index. is superior that of the DJIA index.

1. Correlation analysis for log returns of DJIA and S&P 500 indexes.

By using the formula for correlation, the correlation between the log returns of DJIA and S&P 500 indexes is calculated as follows.

$$\rho_{(DJIA,SP500)} = \frac{\sum_{i=1}^{n}(r_{DJIA} - \bar{r}_{DJIA})(r_{SP500} - \bar{r}_{SP500})}{\sqrt{\sum_{i=1}^{n}(r_{DJIA} - \bar{r}_{DJIA})^2}\sqrt{\sum_{i=1}^{n}(r_{SP500} - \bar{r}_{SP500})^2}}$$

$$where\ r_{DJIA}\ is\ logreturns\ of\ DJIA\ index, r_{SP500}\ is\ logreturns\ of\ S\&P500\ index,$$

$$\bar{r}_{DJIA}\ is\ the\ avergae\ of\ \ r_{DJIA}, \bar{r}_{SP500}\ is\ the\ avarage\ of\ r_{SP500}$$

The correlation $\rho_{(DJIA,SP500)}$ is 0.9653, which implies that the log returns of two indexes are highly correlated with a positive correlation.

2. Mean test for log returns of DJIA and S&P 500 indexes.

To examine whether two samples have equal mean at given significance level, we apply two sample mean t test[2]. The null hypothesis in this two-sample t-test is that two samples have equal mean, while the alternative hypothesis is that two samples do not have equal mean.

Decision rule: Reject H_0 when test statistic lower than or greater than critical value. Otherwise, do not reject H_0.

$$H_0 : \mu_1 = \mu_2 \qquad H_A : \mu_1 \neq \mu_2$$

$$where\ \mu_1\ is\ mean\ of\ log\ returns\ of\ DJIA\ index,$$

$$\mu_2\ is\ mean\ of\ log\ returns\ of\ S\&P500\ index$$

Critical Value: Degree of freedom: v=17501

$$\alpha = 0.05$$

[2] Given that the formula of test statistic is

$$T = \frac{\widehat{\mu_1} - \widehat{\mu_2}}{\sqrt{\frac{\widehat{s_1}^2}{T_1} + \frac{\widehat{s_2}^2}{T_2}}}$$

$$where\ T_1\ and\ T_2\ are\ the\ sample\ sizes, \widehat{\mu_1}\ and\ \widehat{\mu_2}\ are\ the\ sample\ means,$$

$$\widehat{s_1}^2\ and\ \widehat{s_2}^2\ are\ the\ unbaised\ sample\ variance$$

t is the critical value of the t distribution with u degrees of freedom and

$$v = \frac{(\hat{S}_1^2/T_1 - \hat{S}_2^2/T_1)^2}{\sqrt{\frac{(\hat{S}_1^2/T_1)^2}{T_1 - 1} + \frac{(\hat{S}_2^2/T_2)^2}{T_2 - 1}}}$$

$$t_{\frac{\alpha}{2},v} = \pm 1.96$$

Test statistic:

$$t = 0.1434$$

Conclusion: We cannot reject the null hypothesis as the lower critical value <t value <upper critical value. The log returns of DJIA and S&P500 indexes are said to have equal means at the α=5% significance level.

3. Variance equality test log returns of DJIA and S&P 500 indexes.

To examine whether two populations have equal variance at given significance level, we use two-tailed F test[3] which to test whether two populations have equal variance, while the alternative hypothesis is that two populations do not have equal variance.

Decision rule: Reject H_0 when test statistic lower than or greater than critical value. Otherwise, do not reject H_0.

$$H_0 : {\sigma_1}^2 = {\sigma_2}^2 \qquad H_A : {\sigma_1}^2 \neq {\sigma_2}^2$$
$$where\ {\sigma_1}^2\ is\ variance\ of\ logreturns\ of\ DJIA\ index,$$
$$ {\sigma_2}^2\ is\ variance\ of\ logreturns\ of\ S\&P500\ index$$

Critical Value: Degree of freedom: $N_1 = 8756$; $N_2 = 8756$

$$F_{1-\frac{\alpha}{2},N_1-1,N_2-1} = 1.0428 \quad F_{\frac{\alpha}{2},N_1-1,N_2-1} = 0.9590$$

Test statistic:

$$F = 0.9557$$

Conclusion: We reject the null hypothesis as the t value > upper critical value. Based on hypothesis test, the log returns of DJIA and S&P500 indexes do not have equal variances at the α=5% level of significance.

[3] F test statistic is formulated as below:

$$F = \frac{{\widehat{s_1}}^2}{{\widehat{s_2}}^2}$$
$$where\ {\widehat{s_1}}^2\ and\ {\widehat{s_2}}^2\ are\ the\ unbaised\ sample\ variances$$

Part 3 Single Linear Regression (SLR) Analysis of DJIA and S&P 500

The main purpose is to implement a simple linear regression scheme to regress the daily log return of DJIA index r_{it} on the market return, which is the daily log return of S&P 500 r_{mt}. The specification is

$$r_{it} = a + b\, r_{mt} + u_t.$$

We collect both S&P 500 and Dow Jones Index daily closing index from 30 January 1985 to 24 October 2019 and compute the log return for regression (sample size: 8,756). We regress Dow Jones Index as dependent variable and S&P500 as independent variable to observe whether S&P500 can explain the movement of Dow Jones.

So we have the estimation as below:

$$\widehat{r_{it}} = \hat{a} + \hat{b}\, r_{mt} + \widehat{\varepsilon_u}.$$

As the equation above, we estimate the value of $\hat{a}, \hat{b}, \widehat{\varepsilon_u}$.

1. Regression estimator analysis: $\hat{a}$, $\hat{b}$ and their t-stat test.

By using the OLS regression, the coefficients of slope and intercept for the log returns of DJIA on S&P 500 indexes are calculated as follows.

$$\hat{b} = \frac{S_{x,y}}{S_x^2} = 0.9437$$

$$\hat{a} = \bar{y} - \hat{b}\,\bar{x} = 4.2192e^{-5}$$

To examine the significance of $\hat{a}$ *and* $\hat{b}$ within the equation at given significance level, we apply Student's t test[4]:

Decision rule: Reject H_0 when test statistic lower or greater than the critical value. Otherwise, do not reject H_0.

$$H_0 : \hat{a} = 0 \qquad H_1 : \hat{a} \neq 0$$

$$H_0 : \hat{b} = 0 \qquad H_1 : \hat{b} \neq 0$$

Significant Level: $\alpha = 5\%$

Critical Value: Degree of freedom = 8756-1

$$t_{\frac{\alpha}{2},n-1} = \pm 1.96$$

Test statistic: $t_{\hat{a}} = 1.38 \qquad t_{\hat{b}} = 345.95$

Conclusion: $-1.96 < t_{\hat{a}} < 1.96$, therefore, we do not reject H_0 of $\hat{a}$. $\hat{a}$ is not significant

[4] Student's t test is applied to test the significant of the regressor within the equation. The test statistical equation is defined as:

$$t = \frac{\widehat{CE} - 0}{S_{CE}}$$

where S_{CE} *is the the unbaised standard error,*

$$S_{\hat{b}} = \sqrt{\frac{\frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad S_{\hat{a}} = S_{\hat{b}}\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

within the equation. We reject H_0 of $\hat{b}$ as $t_{\hat{b}} > 1.96$. $\hat{b}$ is significant within the equation.

2. Residual analysis: $\widehat{\sigma_u}$

a) Residual Standard Deviation

$$\widehat{\sigma_u} = \sqrt{\frac{RSS}{n-K}} = 0.0029$$

b) Residual distribution analysis: Jarque-Bera test statistic:

To test whether the kurtosis and skewness of the equation match a normal distribution, we introduced Jarque-Bera test:

Decision rule: Reject H0 when test statistic lower or greater than the critical value. Otherwise, do not reject H0.

$$H_0 : JB = 0 \text{ (skewness being zero and the excess kurtosis being zero)}$$
$$H_1 : JB \neq 0 \text{ (At least skewness or kurtosis = 0)}$$
$$\alpha = 0.05$$

Test statistic:

$$JB\ stat = 25350$$

Critical Value: Degree of freedom =2

$$X^2_{0.05,2} = 0.32$$

Conclusion: We reject H_0 as the test statistic > critical value. It indicates that the residual distribution is not normal.

c) Residual auto-correlation analysis: Durbin-Watson statistic:

To test existence of autocorrelation at lag 1 in the residual within the equation, we apply Durbin-Watson test[5]:

Decision rule: Reject H_0 when test statistic lower than d_L or greater than $4 - d_L$. Do not reject H_0 if the test statistic is in between d_U and $4 - d_U$.Otherwise, remain inconclusive.

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$
$$\alpha = 0.05$$

Test statistic:

$$DW\ stat = 1.9058$$

Critical Value: Given n=8756, k= number of regressor excluding intercept.

$$d_{L,(n,k=1)} = 1.664\ , d_{U,(n,k=1)} = 1.684$$

Conclusion: Do not reject H_0 as test statistic falls between d_U and $4 - d_U$.

---

[5] Durbin–Watson statistic is used to detect the presence of autocorrelation at lag 1 in the residuals. the test statistic is:

$$DW = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

d) **Residual distribution analysis: Breusch-Pagan-Godfrey test statistic[6]:**

To test whether all residual have the same finite variance in the equation, we introduced Breusch-Pagan-Godfrey test:

Decision rule: Reject H0 when test statistic lower or greater than the critical value. Otherwise, do not reject H0.

$$H_0 : V_{\varepsilon_i} = \sigma^2 \quad \text{(variance of the error term is constant. (Homoskedasticity))}$$

$$H_1 : V_{\varepsilon_i} \neq \sigma^2 \quad \text{(variance of the error term is not constant. (Heteroskedasticity))}$$

Critical Value:

$$\alpha = 0.05$$

$$X^2_{a,p-1} = X^2_{0.05,1} = 1.96$$

Test statistic:

$$LM\ stat = 256.014$$

Conclusion: We reject H_0 as the test statistic > critical value. It indicates that variance of the error term is not constant at significant level of 5%.

e) **Residual auto-correlation analysis: Breusch-Godfrey Test:**

To test existence of autocorrelation at r lag order in the residual within the equation, we apply Breusch-Godfrey Test[7]:

Decision rule: Reject H_0 when test statistic lower than critical value. Otherwise, do not reject H_0.

$$H_0 : \rho_1 = 0\ and\ \rho_2 = 0\ and \ldots and\ \rho_r = 0$$

$$H_1 : \rho_1 \neq 0\ or\ \rho_2 \neq 0\ or \ldots or\ \rho_2 \neq 0$$

Critical Value

$$X^2_{a,q}$$

Test statistic[8]:

q Lag Order	Chi-Square Critical Value	Test Statistic	P-value
1	1.96	19.24	0.00
2	2.45	22.79	0.00
3	2.80	25.73	0.00
4	3.08	30.80	0.00
5	3.33	33.41	0.00
6	3.55	36.01	0.00
7	3.75	36.57	0.00

[6] It tests whether the variance of the errors from a regression is dependent on the values of the independent variables. In that case, heteroskedasticity is present. We follow three-step procedure to form test statistic: (1) Apply OLS in the model $y = XB + \varepsilon$. (2) Perform auxiliary regression $e_i^2 = \gamma_1 + \gamma_2 z_{2i} + \cdots + \gamma_p z_{pi} + \eta_i$. (3) Test statistic is the production of sample size and R-square.

[7] The Breusch–Godfrey serial correlation LM test is a test for autocorrelation in the errors in a regression model. It makes use of the residuals from the model being considered in a regression analysis, and a test statistic is derived from that. $.u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \cdots + \rho_p u_{t-p} + \varepsilon_t$.

[8] We obtain R^2 from this regression and the test statistic can be show that $(T - r)R^2 \sim \chi^2_r$.

8	3.94	36.57	0.00
9	4.11	38.37	0.00
10	4.28	43.54	0.00

Table 4: Test statistic result and respective $X^2_{a,p}$ critical value.

Conclusion: Reject H_0 from lag 1 to lag 10 correlation at significant level of 5%. The residual do not have the same finite variance in the equation.

3. OLS model reliability analysis: R^2 *and adjusted* R^2:

$$R^2 = \frac{RSS}{ESS} = 0.9318$$

$$\bar{R}^2 = \frac{RSS/(n-K)}{TSS/(n-1)} = 0.9318$$

Conclusions:

- We have Jarque-Bera test result for the residual excess kurtosis or skewness is far from zero. It shows that the **residuals are not normally distributed. As a result, the OLS estimator is not the maximum likelihood estimator**.
- As the **null hypothesis of Jarque-Bera test(normality) is rejected**, the results of t test and two sample t test are said to be invalid.
- Jarque-Bera test is good for testing symmetric distribution with high sample kurosis[9] (Yap and Sim, 2011 However, **Jarque-Bera test is spot to have limitation in test the normality of the sample when the sample size grows, null hypothesis will be rejected, increasingly often, because of the non-stationarity of the data**.
- From result of Student's t test above, we conclude that the **intercept estimator is not significant** within the equation. On the other hand, **the coefficient of S&P500 shows statistical significant in the equation**. However, the assumption of distribution normality is violated.
- Log return of S&P500 has a coefficient of 0.9437, **it proposes that given every one unit of log return of S&P500 increase, the log return of Dow Jones will increase by 0.9437 unit.**
- As our result in table 1.7, we knew that the log return of DJIA and S&P500 both have negative skewness which means the distribution has fat tail or outlier on the negative end. The high kurtosis shows fat tail exist in both DJIA and S&P500 log return.
- Although we conclude that the autocorrelation in residuals does not exist through Durbin Watson test, the assumption of error (residual) is normally distributed in DW test is violated. Then, the **Durbin Watson test is not a valid test in this equation**.
- Limitations of Durbin Watson test include**, it has almost no power except for very high correlations, and it is not valid in presence of lagged dependent variables**[10](Kleiber and Zeileis, 2017).
- We use alternative test to check whether the residual is autocorrelated with first term or higher order term**. Breusch-Godfrey test suggested the residual is autocorrelated at first order**

[9] B. W. Yap & C. H. Sim (2011) Comparisons of various types of normality tests, Journal of Statistical Computation and Simulation, 81:12, 2141-2155, DOI:10.1080/00949655.2010.520163.

[10] Christian Kleiber and Achim Zeileis, 2017, Applied Econometrics With R Chapter 7 Programming, page 13

and higher order terms as well[11].

- We also check the homoskedasticity of the residual, attempting the ensure the assumption of Breusch-Godfrey test is fulfilled. As a result**, the residual of the equation is not homoskedasticity. Then, Breusch-Godfrey test is not valid as well**.
- We obtained a completely different results from Breusch-Godfrey test and Durbin Watson test. Durbin Watson test suggest the autocorrelation is not exist but Breusch-Godfrey test indicate that autocorrelation in residual exist in first or higher order.
- R^2 and $\bar{R}^2$ are both has a same result of 0.9318 as we have only one regressor in the equation causing the $\bar{R}^2$ unable carry out the its meaning of adjusting according the number of regressor. High R^2 and $\bar{R}^2$ indicating **a high linear correlation between the regressand and regressor, and the equation is well explained given the regressor.**
- Given the test conducted above, **we cannot conclude the OLS estimator is the best linear unbiased estimator of the coefficients according to the Gauss–Markov theorem**.

[11] Table 4

QF603 Mini Project

Statistical Analysis of 2 Financial Indexes

An Analysis of Dow Jones Industrial Average (DJIA) and S&P 500

Author:
Team 3

Members:
Wei Sheng Khor
Ruilin He
Zhizhong Zhu
Yimin Peng
Haolin Pan

November 4, 2019

1 Preview of Dow Jones Industrial Average and S&P 500

In this part we analyze two data sets of Dow Jones Industrial Average (DJIA) index and S&P 500 index from **Yahoo! Finance**.

1.1 DJIA index value is equal to the sum of component equity prices divided by devisor.

Symbol	Company Name	Last Price
WMT	Walmart Inc.	119.04
UNH	UnitedHealth Group Incorporated	244.91
HD	The Home Depot, Inc.	234.38
XOM	Exxon Mobil Corporation	69.25
UTX	United Technologies Corporation	142.96
VZ	Verizon Communications Inc.	60.37
MRK	Merck & Co., Inc.	82.26
DIS	The Walt Disney Company	130.9
MSFT	Microsoft Corporation	140.73
NKE	NIKE, Inc.	90.92
JNJ	Johnson & Johnson	128.35
MCD	McDonald's Corporation	194.61
TRV	The Travelers Companies, Inc.	130.43
JPM	JPMorgan Chase & Co.	126.03
CVX	Chevron Corporation	118.67
V	Visa Inc.	177.85
IBM	International Business Machines Corporation	135.44
PFE	Pfizer Inc.	36.77
CSCO	Cisco Systems, Inc.	46.9
AAPL	Apple Inc.	246.58
PG	The Procter & Gamble Company	123.25
BA	The Boeing Company	339.83
GS	The Goldman Sachs Group, Inc.	214.23
KO	The Coca-Cola Company	53.75
AXP	American Express Company	118.26
WBA	Walgreens Boots Alliance, Inc.	55.42
DOW	Dow Inc.	50.48
MMM	3M Company	166.09
CAT	Caterpillar Inc.	139.73
INTC	Intel Corporation	56.46
	Sum of Price	3974.85

Table1: Components of DJI, price as of 2019-10-25 market close

Given the devisor = 0.147445684

$$\begin{aligned} DJIA &= \frac{Sum\,of\,Price}{Devisor} \\ &= \frac{3974.85}{0.147445684} \\ &= 26958.06 \end{aligned}$$

1.2 Analysis

- Due to lagged refreshing frequency, calculating index by one's own may provide advantage in trading.
- The index today consists of 30 U.S. blue-chip stocks. The name "Industrial" is largely historical, as most stocks in this index are not from manufacturing industries, but rather from all the major sectors except utilities and transportation.

1.3 S&P 500 index introduction.

S&P 500 Index Introduction

The S&P 500 or Standard & Poor's 500 Index is a market-capitalization-weighted index of the 500 largest U.S. publicly traded companies. The index is widely regarded as the best gauge of large-cap U.S. equities. And S&P 500 includes some of the components of Dow Jones 30.

Weighting Formula and Calculation for the S&P 500

The S&P 500 uses a market capitalization weighting method, giving a higher percentage allocation to companies with the largest market capitalizations.

$$Company\ Weighting\ in\ S\&P = \frac{Company\ market\ cap}{Total\ of\ all\ market\ caps}$$

Determination of the weighting of each component of the S&P 500 begins with summing the total market cap for the index.

- Calculate the total market cap for the index by adding all the market caps of the individual companies.
- The weighting of each company in the index is calculated by taking the company's market capitalization and dividing it by the total market cap of the index.
- For review, the market capitalization of a company is calculated by taking the current stock price and multiplying it by the company's outstanding shares.
- Fortunately, the total market cap for the S&P as well as the market caps of individual companies is published frequently on financial websites saving investors the need to calculate them.

S&P 500 Index Construction

The market capitalization of a company is calculated by taking the current stock price and multiplying it by the outstanding shares. The S&P only uses free-floating shares, meaning the shares that the public can trade. The S&P adjusts each company's market cap to compensate for new share issues or company mergers. The value of the index is calculated by totaling the adjusted market caps of each company and dividing the result by a divisor. Unfortunately, the divisor is proprietary information of the S&P and is not released to the public. However, we can calculate a company's weighting in the index, which can provide investors with valuable information. If a stock rises or falls, we can get a sense as to whether it might have an impact on the overall index. For example, a company with a 10% weighting will have a greater impact on the value of the index than a company with a 2% weighting.

The Widely Quoted S&P 500

The S&P 500 is one of the most widely quoted American indexes because it represents the largest publicly traded corporations in the U.S. The S&P 500 focuses on the U.S. market's large-cap sector and is also a float-weighted index, meaning company market capitalizations are adjusted by the number of shares available for public trading.

Limitations of the S&P 500 Index

One of the limitations to the S&P and other indexes that are market-cap weighted arises when stocks in the index become overvalued meaning they rise higher than their fundamentals warrant. If a stock has a heavy weighting in the index while being overvalued, the stock typically inflates the overall

value or price of the index. A rising market cap of a company isn't necessarily indicative of a company's fundamentals, but rather it reflects the stock's increase in value relative to shares outstanding. As a result, equal-weighted indexes have become increasingly popular whereby each company's stock price movements have an equal impact on the index.

Key Takeways:

- The DJIA is 30 U.S. stocks picked by the S&P Dow Jones Indices.
- The S&P 500 is 500 U.S. stocks picked by an S&P Dow Jones Indices board.
- The DJIA is calculated through a method of simple mathematical averages.
- The S&P 500 is calculated by giving weights to each stock according to their market value.
- While both indexes are used by investors to determine the general trend of the U.S. stock market, the S&P 500 is more encompassing, as it includes a greater sample of total U.S. stocks.

1.4 Time series of the two indexes

Figure 1.0: DJIA and S&P500 daily closed index from 29/1/1985 to 24/10/2019

Conclusions:

- S&P moves highly consistent with the Dow Jones index in the long run.
- During the major inconsistencies in late 1990s and 2015-2016, S&P 500 reflects the general market better as a result of greater sample of total U.S. stocks.

1.5 Time series of log returns

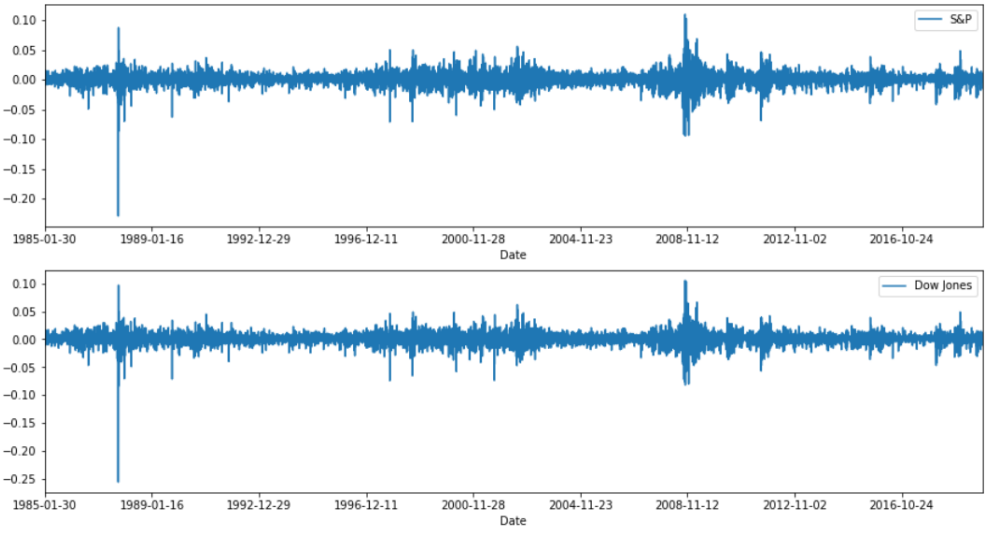


Figure 1.1: Log return of DJIA and S&P500 based on daily closed price from 29/1/1985 to 24/10/2019

1.6 Sample mean and variance of log return

	S&P 500	**Dow Jones**
$\hat{\mu}$	0.032%	0.035%
$\hat{S}^2$	1.259%	1.203%

Table2: Mean and variance of log return of index

1.7 Annualized average and volatility of log return

	S&P 500	**Dow Jones**
Annualized Log return	8.12%	8.73%
Annualized Volatility	17.81%	44.6%

Table3: Annualized log return feature of index

1.8 Sample skewness and sample kurtosis

	S&P 500	**Dow Jones**
Skewness	-1.3	-1.7
Kurtosis	31	44.6

Table4: Skewness and Kurtosis of 2 index

Conclusions:

- The negative skewness for both indexes shows a left skewness, indicating a longer tail in left.
- The high kurtosis shows a feature of fat tail, which predicts movements of three or more standard deviations more frequently than a normal distribution. It's in violation of assumptions in many traditional strategies of asset pricing.

1.9 Jarque-Bera test statistic (JB) and test of normality at the 5% significance level.

To test whether the kurtosis and skewness of the equation match a normal distribution, we use Jarque-Bera test[1]:
Decision rule: Reject H_0 when test statistic lower or greater than the critical value. Otherwise, do not reject H_0.

$$H_0 : JB = 0(\text{skewness being zero and the excess kurtosis being zero})$$

$$H_1 : JB \neq 0(\text{At least skewness or kurtosis= } 0)$$

Significant Level:

$$\alpha = 0.05$$

Critical Value:

$$\chi^2_{(k-1)(r-1)}$$

$$\chi^2_{1,8755} = 31.41$$

Test statistic:

$$JB_{S\&P500} = 287349$$

$$JB_{Dow\ Jones} = 634213$$

Conclusion: We reject the H_0 of both S&P500 and Dow Jones as both of their JB test statistic > critical value.

[1] Jarque–Bera test is a goodness-of-fit test of whether sample data have the skewness and kurtosis matching a normal distribution. The test statistic JB is defined as:

$$JB = \frac{n-k}{6}\left(S^2 + 1/4(K-3)^2\right)$$

where S is the sample skewness, K is the sample kurtosis.
If the data comes from a normal distribution, the JB statistic asymptotically has a chi-squared distribution with two degrees of freedom, so the statistic can be used to test the hypothesis that the data are from a normal distribution. The null hypothesis is a joint hypothesis of the skewness being zero and the excess kurtosis being zero.

2 Log return relationship between DJIA and S&P 500

This part analyzes the relationship between the log returns of DJIA and S&P 500 indexes. In particular, the two-sample t-test is a statistical algorithm to determine if two population means are equal. An application is to test if the log return of S&P 500 index. is superior that of the DJIA index.

2.1 Correlation analysis for log returns of DJIA and S&P 500 indexes.

By using the formula for correlation, the correlation between the log returns of DJIA and S&P 500 indexes is calculated as follows.

$$\rho_{(DJIA,SP500)} = \frac{\sum_{i=1}^{n}(r_{DJIA} - \bar{r}_{DJIA})(r_{SP500} - \bar{r}_{SP500})}{\sqrt{\sum_{i=1}^{n}(r_{DJIA} - \bar{r}_{DJIA})^2}\sqrt{\sum_{i=1}^{n}(r_{SP500} - \bar{r}_{SP500})^2}}$$

where r_{DJIA} is log returns of DJIA index,r_{SP500} is log returns of S&P500 index, r_{DJIA} is the avergae of $\bar{r}_{DJIA}$,r_{SP500} is the avarage of $\bar{r}_{SP500}$

The correlation $\rho_{(DJIA,SP500)}$ is 0.9653, which implies that the log returns of two indexes are highly correlated with a positive correlation.

2.2 Mean test for log returns of DJIA and S&P 500 indexes.

To examine whether two samples have equal mean at given significance level, we apply two sample mean t test [2] . The null hypothesis in this two-sample t-test is that two samples have equal mean, while the alternative hypothesis is that two samples do not have equal mean.

Decision rule: Reject H_0 when test statistic lower than or greater than critical value. Otherwise, do not reject H_0.

$$H_0 : \mu_1 = \mu_2, H_A : \mu_1 \neq \mu_2$$

where μ_1 is mean of log returns of DJIA index, μ_2 is mean of log returns of S&P500 index

Critical Value: Degree of freedom: $v = 17501$

$$\alpha = 0.05$$

$$t_{\frac{\alpha}{2},v} = \pm 1.96$$

Test statistic:

$$t = 0.1434$$

Conclusion: We cannot reject the null hypothesis as the lower critical value <t value <upper critical value. The log returns of DJIA and S&P500 indexes are said to have equal means at the $\alpha = 5\%$ significance level.

[2] Given that the formula of test statistic is

$$T = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{\hat{S_1}^2}{T_1} + \frac{\hat{S_2}^2}{T_2}}}$$

where T_1 and T_2 are the sample sizes,
$\hat{\mu}_1$ and $\hat{\mu}_2$ are the sample means,
$\hat{S_1}^2$ and $\hat{S_2}^2$ are the unbaised sample variance.
t is the critical value of the t distribution with u degrees of freedom and

$$v = \frac{(\hat{S_1}^2/T_1 - \hat{S_2}^2/T_2)^2}{\sqrt{\frac{(\hat{S_1}^2/T_1)^2}{T_1 - 1} - \frac{(\hat{S_2}^2/T_2)^2}{T_2 - 1}}}$$

2.3 3. Variance equality test log returns of DJIA and S&P 500 indexes.

To examine whether two populations have equal variance at given significance level, we use two-tailed F test which to test whether two populations have equal variance, while the alternative hypothesis is that two populations do not have equal variance.
Decision rule: Reject H_0 when test statistic lower than or greater than critical value. Otherwise, do not reject H_0.

$$H_0 : \sigma_1^2 = \sigma_2^2, H_A : \sigma_1^2 \neq \sigma_2^2$$

where σ_1^2 is variance of logreturns of DJIA index, σ_2^2 is variance of logreturns of S&P500 index
Critical Value: Degree of freedom: $N_1 = 8756$; $N_2 = 8756$

$$F_{1-\frac{\alpha}{2},N_1-1,N_2-1} = 1.0428$$

$$F_{\frac{\alpha}{2},N_1-1,N_2-1} = 0.9590$$

Test statistic[3]:

$$F = 0.9557$$

Conclusion: We reject the null hypothesis as the t value > upper critical value. Based on hypothesis test, the log returns of DJIA and S&P500 indexes do not have equal variances at the $\alpha = 5\%$ level of significance.

[3] F test statistic is formulated as below:

$$F = \frac{\hat{S_1}^2}{\hat{S_2}^2}$$

where S_1^2 and S_2^2 are the unbaised sample variances.

3 Single Linear Regression (SLR) Analysis of DJIA and S&P 500

The main purpose is to implement a simple linear regression scheme to regress the daily log return of DJIA index r_{it} on the market return, which is the daily log return of S&P 500 r_{mt} . The specification is

$$r_{it} = a + br_{mt} + u_t$$

We collect both S&P 500 and Dow Jones Index daily closing index from 30 January 1985 to 24 October 2019 and compute the log return for regression(sample size: 8,756). We regress Dow Jones Index as dependent variable and S&P500 as independent variable to observe whether S&P500 can explain the movement of Dow Jones. So we have the estimation as below:

$$\hat{r}_{it} = \hat{a} + \hat{b}r_{mt} + \varepsilon_u$$

As the equation above, we estimate the value of $\hat{a}, \hat{b}, \hat{\varepsilon}_u$.

3.1 Regression estimator analysis: $\hat{a}$ $\hat{b}$ and their t-stat test.

3.1.1 Estimators:

$$\hat{b} = \frac{S_{x,y}}{S_x^2} = 0.9437$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 4.2192e^{-5}$$

3.1.2 Student T test:

To test the significance of $\hat{a}$ and $\hat{b}$ within the equation, we use student t test[4]
Decision rule: Reject H_0 when test statistic lower or greater than the critical value. Otherwise, do not reject H_0.

$$H_0 : \hat{a} = 0, H_1 : \hat{a} \neq 0$$

$$H_0 : \hat{b} = 0, H_1 : \hat{b} \neq 0$$

Significant Level:

$$\alpha = 0.05$$

Critical Value: Degree of freedom = 8756-1

$$t_{\frac{\alpha}{2},n-1} = t_{0.975,8755} = \pm 1.96$$

Test Statistic:

$$t_{\hat{a}} = 1.38$$

$$t_{\hat{b}} = 345.959$$

Conclusion:$-1.96 < t_{\hat{a}} < 1.96$, therefore, we do not reject H_0 of $\hat{a}$. $\hat{a}$ is not significant within the equation at significant level of 5%. We reject H_0 of $\hat{b}$ as $t_{\hat{b}} > 1.96$. $\hat{b}$ is significant within the equation at significant level of 5%.

[4]Student's t test is applied to test the significant of the regressor within the equation. The test statistical equation is defined as:

$$t = \frac{\hat{CE} - 0}{S_{CE}}$$

where S_{CE} is the the unbaised standard error,

$$S_{\hat{b}} = \sqrt{\frac{\frac{1}{n-1}\sum_{i=1}^{n}\hat{\varepsilon}_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \; S_{\hat{a}} = S_{\hat{b}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2}$$

3.2 Residual analysis: $\hat{\sigma_u}$

$$\hat{\sigma}_u = \frac{RSS}{n-K} = 0.00000820159$$

3.2.1 Residual distribution analysis: Jarque-Bera test statistic

To test whether the kurtosis and skewness of the equation match a normal distribution, we apply Jarque-Bera test.
Decision rule: Reject H_0 when test statistic lower or greater than the critical value. Otherwise, do not reject H_0.

$$H_0 : JB = 0 \text{ (skewness and excess kurtosis being zero)}$$

$$H_1 : JB \neq 0 \text{ (At least skewness or kurtosis not being 0)}$$

Critical Value:

$$\alpha = 0.05$$

$$\chi^2_{\alpha,(c-1)(r-1)} = \chi^2_{0.05,8755} = 92.40$$

Test statistic:

$$JB_u = 25350$$

Conclusion: We reject H_0 as the test statistic > critical value. It indicates that the residual is not normally distributed at significant level of 5%.

3.2.2 Residual auto-correlation analysis: Durbin-Watson statistic

To test existence of autocorrelation at lag 1 in the residual within the equation, we apply Durbin-Watson test[5]:
Decision rule: Reject H_0 when test statistic lower than d_L or greater than 4-d_L. Do not reject H_0 if the test statistic is in between d_U and 4-d_U.Otherwise, remain inconclusive.

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

Critical Value: Given n= 8756, k= number of regressor excluding intercept

$$d_{L(n,k=1)} = 1.664 \; d_{U(n,k=1)} = 1.684$$

Test statistic:

$$DW = 1.9058$$

Conclusion: Do not reject H_0 as test statistic falls between d_U and 4-d_U. no frist order correlation exist in the residual.

[5]Durbin–Watson statistic is used to detect the presence of autocorrelation at lag 1 in the residuals. Null hypothesis of DW test: $\rho = 0$, alternative hypothesis $\rho \neq 0$, then the test statistic is

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2)}{\sum_{t=1}^{T} e_t^2}$$

3.2.3 Residual distribution analysis: Breusch-Pagan-Godfrey test statistic

To test whether all residual have the same finite variance in the equation, we introduced Breusch-Pagan-Godfrey test[6]:.
Decision rule: Reject H_0 when test statistic lower or greater than the critical value. Otherwise, do not reject H_0.

$$H_0 : V_{\varepsilon_j} = \sigma^2 \text{ (variance of the error term is constant. (Homoskedasticity))}$$
$$H_1 : V_{\varepsilon_j} \neq \sigma^2 \text{ (variance of the error term is not constant. (Heteroskedasticity))}$$

Critical Value:

$$\alpha = 0.05$$
$$\chi_{\alpha,(p-1)^2}=\chi^2_{0.05,1}=1.96$$

Test statistic:

$$LM\ stat = 256.014$$

Conclusion: We reject H_0 as the test statistic > critical value. It indicates that variance of the error term is not constant at significant level of 5%.

3.2.4 Residual auto-correlation analysis: Breusch-Godfrey Test

To test existence of autocorrelation at r lag order in the residual within the equation, we apply Breusch-Godfrey Test[7]
Decision rule: Reject H_0 when test statistic lower than d_L or greater than 4-d_L. Do not reject H_0 if the test statistic is in between d_U and 4-d_U.Otherwise, remain inconclusive.

$$H_0 : \rho_1 = 0\ and\ \rho_2 = 0\ and \ldots and\ \rho_r = 0$$
$$H_1 : \rho_1 \neq 0\ or\ \rho_2 \neq 0\ or \ldots or\ \rho_r \neq 0$$

Critical Value:

$$\chi^2_q$$

Test statistic[8]:

q lag order	χ^2 Critical Value	Test Statistic	P value
1	1.96	19.24	0.00
2	2.45	22.79	0.00
3	2.80	25.73	0.00
4	3.08	30.80	0.00
5	3.33	33.41	0.00
6	3.55	36.01	0.00
7	3.75	6.57	0.00
8	3.94	36.57	0.00
9	4.11	38.37	0.00
10	4.28	43.54	0.00

Table 4: Test statistic result and respective χ^2 critical value.

Conclusion: Reject H_0 as test statistic more than critical value at significant level of 5%. The residual is heteroskedasticity in the equation.

[6]It tests whether the variance of the errors from a regression is dependent on the values of the independent variables. In that case, heteroskedasticity is present. We follow three-step procedure to form test statistic: (1) Apply OLS in the model $y = XB+\varepsilon$. (2) Perform auxiliary regression $e_i^2 = \gamma_1+\gamma_2 z_{2i}+\cdots+\gamma_p z_{pi}+\eta_i$. (3) Test statistic is the production of sample size and R-square.

[7]The Breusch–Godfrey serial correlation LM test is a test for autocorrelation in the errors in a regression model. It makes use of the residuals from the model being considered in a regression analysis, and a test statistic is derived from that. $u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \cdots + \rho_p u_{t-p} + \varepsilon_t$.

[8]We obtain R^2 from residual regression and the test statistic can be show that $(T-r)R^2 \sim \chi^2_r$.

3.3 OLS model reliability analysis: R^2 and $\bar{R}^2$

$$R^2 = \frac{RSS}{TSS} = 0.9318$$

$$\bar{R}^2 = \frac{RSS/(n-K)}{TSS/(n-1)} = 0.9318$$

3.4 Conclusions

- We have Jarque-Bera test result for the residual excess kurtosis or skewness is far from zero. It shows that the residuals are not normally distributed. As a result, the OLS estimator is not the maximum likelihood estimator.
- As the null hypothesis of Jarque-Bera test(normality) is rejected, the results of t test and two sample t test are said to be invalid.
- Jarque-Bera test is good for testing symmetric distribution with high sample kurosis (Yap and Sim, 2011) [9] . However, Jarque-Bera test is spot to have limitation in test the normality of the sample when the sample size grows, null hypothesis will be rejected, increasingly often, because of the non-stationarity of the data.
- From result of Student's t test above, we conclude that the intercept estimator is not significant within the equation. On the other hand, the coefficient of S&P500 shows statistical significant in the equation. However, the assumption of distribution normality is violated.
- Log return of S&P500 has a coefficient of $\hat{b} = 0.9437$, it proposes that given every one unit of log return of S&P500 increase, the log return of Dow Jones will increase by 0.9437 unit.
- As our result in table 1.7, we knew that the log return of DJIA and S&P500 both have negative skewness which means the distribution has fat tail or outlier on the negative end. The high kurtosis shows fat tail exist in both DJIA and S&P500 log return.
- Although we conclude that the autocorrelation in residuals does not exist through Durbin Watson test, the assumption of error (residual) is normally distributed in DW test is violated. Then, the Durbin Watson test is not a valid test in this equation.
- Limitations of Durbin Watson test include, it has almost no power except for very high correlations, and it is not valid in presence of lagged dependent variables(Kleiber and Zeileis, 2017) [10].
- We use alternative test to check whether the residual is autocorrelated with first term or higher order term. Breusch-Godfrey test suggested the residual is autocorrelated at first order and higher order terms as well [11].
- We also check the homoskedasticity of the residual, attempting the ensure the assumption of Breusch-Godfrey test is fulfilled. As a result, the residual of the equation is not homoskedasticity. Then, Breusch-Godfrey test is not valid as well.
- We obtained a completely different results from Breusch-Godfrey test and Durbin Watson test. Durbin Watson test suggest the autocorrelation is not exist but Breusch-Godfrey test indicate that autocorrelation in residual exist in first or higher order.

[9] B. W. Yap & C. H. Sim (2011) Comparisons of various types of normality tests, Journal of Statistical Computation and Simulation, 81:12, 2141-2155, DOI:10.1080/00949655.2010.520163

[10] Christian Kleiber and Achim Zeileis, 2017, Applied Econometrics With R Chapter 7 Programming, page 13

[11] Table 4

- R^2 and $\bar{R}^2$ are both has a same result of 0.9318 as we have only one regressor in the equation causing the $\bar{R}^2$ unable carry out its function of adjusting according the number of regressor. From R^2 and $\bar{R}^2$, approximate 93.18% of the variation in the output variable is explained by the input variable. It also indicates a high linear correlation between the regressand and regressor.
- Given the test conducted above, we cannot conclude the OLS estimator is the best linear unbiased estimator of the coefficients according to the Gauss–Markov theorem.

4 Appendix Code

Listing 1: Python Code

```
from __future__ import print_function, division
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import scipy.stats as stats

#1.3 Plot time series of two indexes
DJIA = pd.read_csv('^DJI.csv')
DJIA.index = pd.to_datetime(DJIA.Date)
plt.plot(DJIA.Close)
plt.show()

SP500 = pd.read_csv('^GSPC.csv')
SP500.index = pd.to_datetime(SP500.Date)
plt.plot(SP500.Close)
plt.show()

#1.4.1 Daily log return on DJIA index
Rt_DJIA = pd.DataFrame(DJIA.Close)
Rt_DJIA['log_Rt'] = np.log(Rt_DJIA.Close)-np.log(Rt_DJIA.Close.shift(1))
print(Rt_DJIA)

#1.4.2 Daily log return on S&P500 index
Rt_SP500 = pd.DataFrame(SP500.Close)
Rt_SP500['log_Rt'] = np.log(Rt_SP500.Close)-np.log(Rt_SP500.Close.shift(1))
print(Rt_SP500)

#1.5 Plot the time series of log return on DJIA and SP500 index
DJIA['log_Rt'] = np.log(Rt_DJIA.Close)-np.log(Rt_DJIA.Close.shift(1))
print(DJIA)
SP500['log_Rt'] = np.log(Rt_SP500.Close)-np.log(Rt_SP500.Close.shift(1))
print(SP500)
plt.plot(DJIA.log_Rt)
plt.show()
plt.plot(SP500.log_Rt)
plt.show()

#1.6 Sample mean and unbaised smaple variance of log returns
Mean_DJIA = DJIA.log_Rt.mean()
Var_DJIA = DJIA.log_Rt.var()
Std_DJIA = DJIA.log_Rt.std()
print('The sample mean of log return on DJIA is {:.4}'.format(Mean_DJIA) +
      ', ' + 'the unbaised sample variance is {:.4}.'.format(Var_DJIA))

Mean_SP500 = SP500.log_Rt.mean()
Var_SP500 = SP500.log_Rt.var()
Std_SP500 = SP500.log_Rt.std()
print('The sample mean of log return on S&P500 is {:.4}'.format(Mean_SP500) +
      ', ' + 'the unbaised sample variance is {:.4}.'.format(Var_SP500))

#1.7 Annualized average and volatility of log return in percent, each year has 252 trading days
Ann_mean_DJIA = 252 * Mean_DJIA
Ann_mean_SP500 = 252 * Mean_SP500
Ann_std_DJIA = np.sqrt(252 * Std_DJIA**2)
Ann_std_SP500 = np.sqrt(252 * Std_SP500**2)
print('The annualized average of log return on DJIA in percent is {:.2%}'.format(Ann_mean_DJIA) +
', ' + 'the annualized volatility in percent is {:.2%}.'.format(Ann_std_DJIA))
print('The annualized average of log return on S&P500 in percent is {:.2%}'.format(Ann_mean_SP500) +
', ' + 'the annualized volatility in percent is {:.2%}.'.format(Ann_std_SP500))

#1.8 Sample skewness and sample kurtosis
def skewkurt(m):
    s = 0
```

```
    for i in range(1,len(m)):
        s = s + m[i]
    mu = s/(len(m)-1)

    s = 0
    for i in range(1,len(m)):
        s = s + (m[i] - mu)**2
    var = s/(len(m)-1)

    std = np.sqrt(var)
    s = 0
    for i in range(1,len(m)):
        s = s + (m[i] - mu)**3
    skew = s/((len(m)-1)*std**3)
    sample_skew = round(skew,4)

    s = 0
    for i in range(1,len(m)):
        s = s + (m[i] - mu)**4
    kurt = s/((len(m)-1)*std**4)
    sample_kurt = round(kurt,4)

    return ('the sample skewness is ' + str(sample_skew) + ', ' + 'the sample kurtosis is ' +
     str(sample_kurt) + '.')

print('For DJIA index, ' + skewkurt(DJIA.log_Rt))
print('For S&P500 index, ' + skewkurt(SP500.log_Rt))

#1.9 Jarque-Bera test statistic
JBtest_DJIA = stats.jarque_bera(DJIA.log_Rt[1:])
JBtest_SP500 = stats.jarque_bera(SP500.log_Rt[1:])

def hypothesis(x):
    if x[1]>0.05:
        print('We cannot reject the null hypothesis H0: JB=0')
    else:
        print('We reject the null hypothesis H0: JB=0, ' + 'the Jarque-Bera test statistic is '
        + str(round(x[0],4)))

hypothesis(JBtest_DJIA)
hypothesis(JBtest_SP500)

#2.1 Correlation between log returns of DJIA and S&P500 indexs
Corr_DJSP = np.corrcoef(DJIA.log_Rt[1:],SP500.log_Rt[1:])
print('The correlation between the log returns of DJIA and S&P500 indexes is ' +
str(round(Corr_DJSP[0][1],4)) + '.')

#2.2 Examine whether 2 samples have equal mean at the alpha = 5% significance level
Mu1 = Mean_DJIA
Mu2 = Mean_SP500
Var1 = Var_DJIA
Var2 = Var_SP500
T1 = len(DJIA.log_Rt)-1
T2 = len(SP500.log_Rt)-1
T_stats = (Mu1-Mu2)/np.sqrt(Var1/T1+Var2/T2)
Dof = (Var1/T1+Var2/T2)**2/((Var1/T1)**2/(T1-1)+(Var2/T2)**2/(T2-1))
alpha = 0.05

if np.abs(T_stats) > stats.t.ppf(1-alpha/2,Dof):
    print('We reject the null hypothesis H0: mu_DJI=mu_SP500, and accept the alternative
    hypothesis Ha: mu_DJImu_SP500')
elif T_stats > stats.t.ppf(1-alpha,Dof):
    print('We reject the null hypothesis H0: mu_DJI=mu_SP500, and accept the alternative
    hypothesis Ha: mu_DJI>mu_SP500')
elif T_stats < stats.t.ppf(alpha,Dof):
    print('We reject the null hypothesis H0: mu_DJI=mu_SP500, and accept the alternative
    hypothesis Ha: mu_DJI<mu_SP500')
```

```
else:
    print('We can not reject the null hypothesis H0: mu_DJI=mu_SP500')

#2.3 F-test for equality of 2 variances at the alpha=5% level of significance
F_stats = Var1/Var2
Dof1 = T1-1
Dof2 = T2-1
alpha = 0.05

if F_stats < stats.f.ppf(1-alpha/2,Dof1,Dof2) or F_stats > stats.f.ppf(alpha/2,Dof1,Dof2):
    print('We reject the null hypothesis H0: sigma_DJI=sigma_SP500, and accept the
    alternative hypothesis Ha: sigma_DJIsigma_SP500')
else:
    print('We can not reject the null hypothesis H0: sigma_DJI=sigma_SP500')

from scipy.stats import linregress as ols
class stat_SLR:
    def __init__(self,x,y):
        x=np.array(x)
        y=np.array(y)
        self.b, self.a, *other = ols(x,y)
        self.T = len(x)
        n=self.T
        # sample property
        self.x_mean = x.mean()
        self.y_mean = y.mean()
        self.x_sigma = x.std(ddof=1)
        self.y_sigma = y.std(ddof=1)
        self.y_hat = self.b*x+self.a
        self.u_hat = y-self.y_hat
        # prediction check
        self.RSS = ((self.u_hat)**2).sum()
        self.TSS = ((y-self.y_mean)**2).sum()
        self.ESS = self.TSS - self.RSS
        self.R_2 = 1-self.RSS/self.TSS
        self.R_2_adj = 1- self.RSS/(self.T-2)/(self.TSS/(self.T-1))
        # estimator check
        self.u_sigma_hat = np.sqrt(self.RSS/(self.T-2))
        self.b_sigma = self.u_sigma_hat/np.sqrt(x.std(ddof=0)**2*n)
        self.a_sigma = self.b_sigma*np.sqrt(x.std(ddof=0)**2+x.mean()**2)
        # ddof = T-2
        self.a_hat_tstat = self.a/self.a_sigma
        self.b_hat_tstat = self.b/self.b_sigma

    @staticmethod
    def DW_test(s):
        return ((s-np.append(np.nan,s[:-1]))**2)[1:].sum()/(s**2).sum()


t3 = stat_SLR(log_r.SP500,log_r.DJI)
slope, intercept = t3.b, t3.a
print('a={}, b={}'.format(intercept, slope))
sigma_u = t3.u_sigma_hat
print('sigma_u = {}'.format(sigma_u))
a_hat_t = t3.a_hat_tstat
b_hat_t = t3.b_hat_tstat
print('t stat for a = {}, for b = {}'.format(a_hat_t, b_hat_t))

get_critical = lambda q, dof: ss.t.ppf(q, df=dof)
dof = t3.T-2
q=1-0.05/2
critical = get_critical(q,dof)
print('Critical value for a and b is {}'.format(critical))

R_2 = t3.R_2
R_2_adj = t3.R_2_adj
print('R square = {}, adjusted R square = {}'.format(R_2, R_2_adj))
```

```
JB_u = JB_test(t3.u_hat)
print('JB test result is {}'.format(JB_u))

DW_u = stat_SLR.DW_test(t3.u_hat)
print('DW test result is {}'.format(DW_u))
```