STAT-1043 Statistical Techniques and Time Series

Name: Yogi Bhoot

ID: 001189309

Date: 07/12/2021

I. Obtain the time series using the correction code syntax making use of the code below.

getSymbols("^GSPC", from=as.Date("2009-01-01"), to=as.Date("22020-12-31"), periodicity="daily")

The getSymbol function is accessed by the quantmod package. getSymbol function used for fetching data from different sources like online, local and others. Here, we get daily periodic GSPC(S&P500) data between 01/01/2009 to 31/12/2020 from Yahoo Finance website.

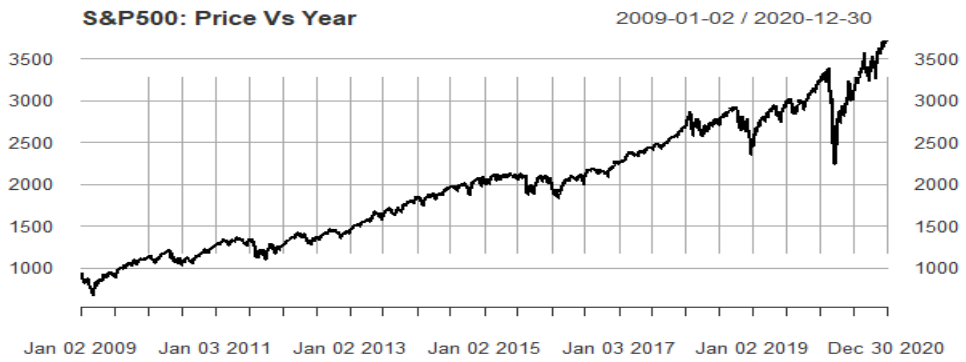


Figure 1. S&P500 Time Series

II. Transfer your series to log return.

ret = diff(log(price_adj))

Equation**:-** $\mathbf{rt = ln\left(\frac{(Pt)}{(Pt-1)}\right) = ln(Pt) - ln(Pt - 1)}$

Log return is used to remove trend and seasonality which gives us stationary time series by differentiating the current time(t) from previous time(t_1).

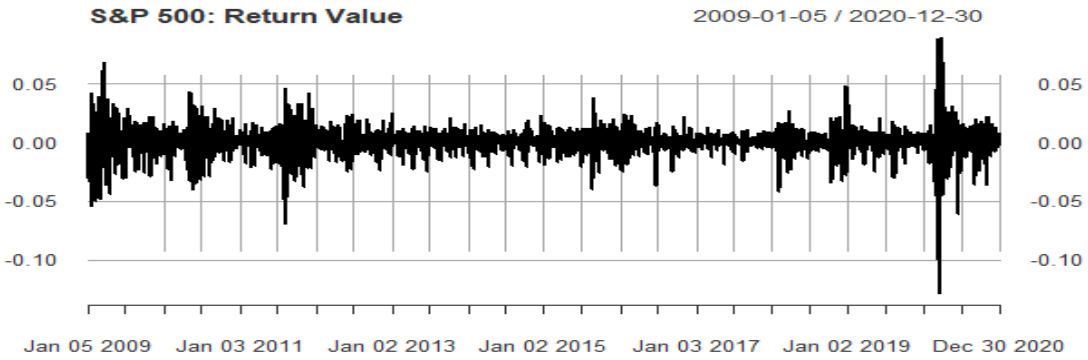


Figure 2. S&P500 log return Time Series

III. Examine the ACF and PACF functions.

acf(retn, main = "Autocorrelation for S&P500")
pacf(retfn, main = "Partial Autocorrelation for S&P500")

ACF test measures direct and indirect relation between observations of time series that are divided by k time stamp. PACF test describe that dependence between an observation and its lag. We use ACF and PACF function to determine which model is best for time series according their significant values that means lags value. Here, we observed MA4 and AR4 model is best fit for our time series.

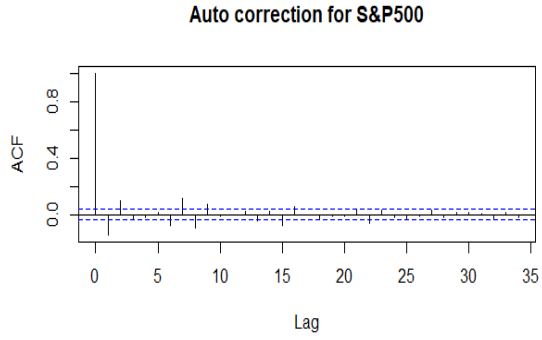


Figure 3. ACF TEST

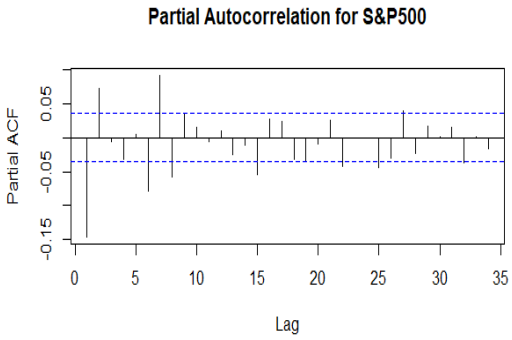


Figure 4. ACF TEST

IV. Perform the Ljung-Box test and describe the test-hypothesis and report / comment on the result.

Box.test(retn, type = "Ljung-Box")

Equation**:-** $\mathbf{Q(m) = n(n+2)\sum_{j=1}^{m} \frac{r_j^2}{(n-j)}}$

Where, r_j = sample autocorrelation, m = time lag.

Ljung-Box test is used for checking whether data is auto-correlation in series. If output p value is smaller than 0.05, indicate the data series is considerable auto-correlation. If the value of p is bigger than 0.05 that means it rejected null hypothesis. After perform test, we got p value less than 0.05 which consider our time series is auto-correlation.

O/p:- X-squared = 64.8, df = 1, p-value = 7.772e-16

V. Check the data for stationary using the correct the statistic and comment on the output

adf.test(retn, alternative = c("stationary"))

Equation**:-** $\mathbf{y_t = c + \beta t + \alpha y_{t-1} + \emptyset \Delta Y_{t-1} + e_t}$

The Augmented Dickey Fuller Test is unit root test for checking time series is stationary or not. ADF is more complex model compare to DF Test. If p value is lower than critical or significant value (0.05) which means series is stationary. Also, it know null hypothesis. Alternative hypothesis is differ slightly related its equation. But, it use for trend stationary series. Here, we test ADF test for checking our series is stationary or

not. We found p value is 0.01 smaller than 0.05 that means our series is stationary. Also, alternative hypothesis is stationary.

O/p:- Dickey-Fuller = -15.327, Lag order = 14, p-value = 0.01
Alternative hypothesis: stationary

VI. Perform a normality test of your choice on the return series and report the output. Write down the hypothesis test and comment on the p-value

shapiro.test(as.vector(retn))

Equation:-W $= \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

Shapiro test use for checking normality test on time series. According theory, if p value is smaller than alpha value that means test rejected null hypothesis for normal distribution. If alpha value is less than p value which is our series normal distribution. In our case, our value is 2.2e-16 that means our series considerable for normal distribution.

O/p:- W = 0.87502, p-value < 2.2e-16

VII. Fit an ARIMA model and determine the correct lag order: Show the 1-liner codes for output.
armodel = auto.arima(retn, trace = TRUE)

ARIMA model is a statistical analysis model that is used to predict future value based on historical data. In the ARIMA model, AR (Autoregression) shows a changing variable which means regresses on values. I (Integrated) shows the differencing between current time data and past time data to make stationary series. MA (Moving average) shows the dependency between an observation and residual error from the MA model applied to lag. All parameters in the ARIMA function have standard notation. standard notation is p, d and q, where integer values for parameters show what type of ARIMA model to use. P is Lag order. D is the degree of differencing. Q is size of moving average. Here, we used an auto.arima model to fit our time series that gives the right order model. We found the best ARIMA model is (4,0,4) for our time series. Where p = 4, d = 0 and q =4.AIC is estimate of prediction error and quality of model for exist data set. BIC is measurement of selection model for finite data set model.

$$\boldsymbol{AIC = -2\log(L) + 2(p+q+k)}$$

$$\boldsymbol{AICc = AIC + \frac{2(p+q+k)(p+q+k+1)}{T-p-q-k-1}}$$

$$\boldsymbol{BIC = AIC + \big((\log T) - 2\big)(p+q+k)}$$

Where, k is intercept of the ARIMA model.
O/p:- ARIMA(4,0,4) with zero mean
sigma^2 estimated as 0.0001306: log likelihood=9220.14
AIC=-18422.29 AICc=-18422.23 BIC=-18368.17

VIII. Report the coefficients for the chosen ARIMA model and show the respective equation given these coefficients.

armodel$coef

MA4 = $\boldsymbol{\mu + w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \theta_3 w_{t-3} + \theta_4 w_{t-4}}$
AR4 = $\boldsymbol{\beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \beta_3 y_{t-3} + \beta_4 y_{t-4} + \epsilon_t}$

$$\boldsymbol{y_t = m + 0.8 y_{t-1} + 0.1 y_{t-2} + e_t}$$

Here, we got coefficient by perform above equation. Time series perform upto MA4 and AR4 equation

O/p:-

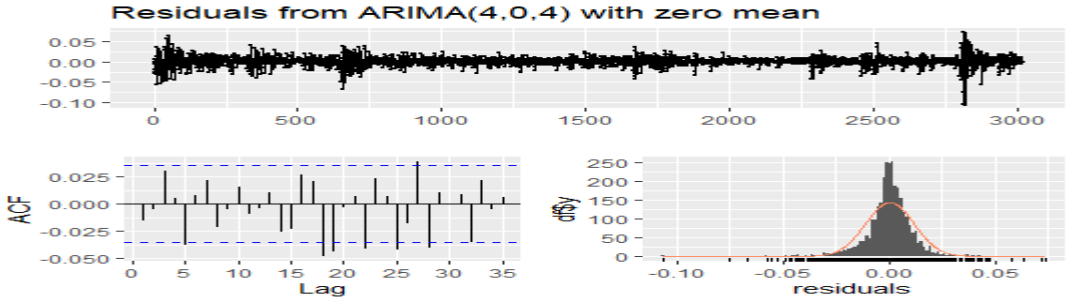
ar1	ar2	ar3	ar4	ma1	ma2	ma3	ma4
-0.11216 5711	0.81977 8916	-0.28478 9007	-0.83146 6335	0.00564 1687	-0.75110 0035	0.32911 6269	0.71741 8126

IX. The residuals from an ARIMA fit require that:

a. The residuals have zero mean $E[e_E] = 0$
b. Have a finite variance $Var\mathrm{E}[e_E]$=0
c. Have zero autocovariance $\mathrm{E}[e_t e_\tau] = 0$

checkresiduals(armodel)

The autocorrelation function show no significant between the residual.
The p value is 0.00749 which means Ljung-Box shows making error rate is very low.
The graph is well shaped and suggesting they are nearly symmetric.



O/p:- Ljung-Box test
data: Residuals from ARIMA(4,0,4) with zero mean
Q* = 11.969, df = 3, p-value = 0.00749

Model df: 8. Total lags used: 11